Is Britney Spears Spam?

Aaron Zinman
Sociable Media Group
MIT Media Lab
Cambridge, MA 02139
+1.617.452.5606
azinman@media.mit.edu

Judith Donath
Sociable Media Group
MIT Media Lab
Cambridge, MA 02139
+1.617.253.5098
judith@media.mit.edu

ABSTRACT

We seek to redefine spam and the role of the spam filter in the context of Social Networking Services (SNS). SNS, such as MySpace and Facebook, are increasing in popularity. They enable and encourage users to communicate with previously unknown network members on an unprecedented scale. The problem we address with our work is that users of these sites risk being overwhelmed with unsolicited communications not just from e-mail spammers, but also from a large pool of well intending, yet subjectively uninteresting people.

Those who wish to remain open to meeting new people must spend a large amount of time estimating deception and utility in unknown contacts. Our goal is to assist the user in making these determinations. This requires identifying clear cases of undesirable spam and helping them to assess the more ambiguous ones. Our approach is to present an analysis of the salient features of the sender's profile and network that contains otherwise hard to perceive cues about their likely intentions.

As with traditional spam analysis, much of our work focuses on detecting deception: finding profiles that mimic ordinary users but which are actually commercial and usually undesirable entities. We address this within the larger context of making more legible the key cues presented by any unknown contact.

We have developed a research prototype that categorizes senders into broader categories than spam/not spam using features unique to SNS. We discuss our initial experiment, and its results and implications.

1. INTRODUCTION

People use social networking services (SNS) such as MySpace and Friendster both to stay in touch with people in their existing social network and to expand their networks by establishing new connections. Communication with strangers is often an inherent part of that world: SNS exist in part to enable unsolicited, yet friendly and welcome communication.

This openness to messages from strangers leaves users of these sites vulnerable to a growing quantity of unwelcome contact, including spam. Some would look familiar to any email user: ads for Viagra™ and breathy invitations to pornographic websites. Some have agendas that are more ambiguous: is that "friend request" from an attractive stranger a genuine gesture from someone intrigued by your witty profile, or is it phony façade that will lead to a torrent of advertising? And others are a different class of unwelcome communication: contact from an unknown person that is not malicious or deceptive, yet is still unwanted by the recipient. Differentiating between welcome and unwelcome communication is subjective, dependent on the taste and interests of the user.

The problem we address with this work is that users of these sites risk being overwhelmed with unsolicited communications. Those who wish to remain open to meeting new people must spend a large amount of time evaluating these contacts to determine which are desirable. Our goal is to assist the user in making this determination. Identifying clear cases of undesirable spam is one part of the task, but it also requires helping them to assess the more ambiguous ones. Our approach is to present an analysis of the salient features of the sender's profile and network that provides useful and otherwise hard to perceive cues about their likely intentions.

Understanding the culture of SNS is essential for this task. Their key features are that each user is represented by a self-made descriptive profile that includes links to other chosen members of the site, and that users can leave publicly readable comments on their friends' profiles. These links to other users make up the extensive network that characterizes these sites.

Links on most sites are mutual, i.e. both parties must agree to connect to each other. To reduce spam, many sites (including MySpace, the subject of our analysis) require that one be a member of someone's personal network of links in order to leave comments on their profile. Spammers often pose as attractive young girls or other intriguing characters to lure users to accept them into their personal network. Once a member, they exploit the connection by flooding the unwary user with a barrage of unexpected advertisements. To prevent such attacks, users must first judge the character of a virtual identity using any available information before accepting the connection. As the number of friend requests increases, so does the cost of manually examining each profile.

With the continued growth of social networking comes greater promotional use of these sites. Entities from pornographic websites to political candidates are attracted to the huge audience and atmosphere of trust. While some users are not interested in connecting with any such commercial groups, others are. They may welcome news about a favorite band's upcoming tour; they may wish to express their liking for a celebrity, a political candidate, or a brand of beer. Many even enjoy populating their personal network with pornographic profiles.

CEAS 2007 – Fourth Conference on Email and Anti-Spam, August 2-3, 2007, Mountain View, California USA

The definition of what constitutes spam in SNS is thus subjective. For example, one might receive a friend request from a celebrity such as Britney Spears[1]. Whether we like her or her music may be independent of whether we wish to interact with her virtual persona. But unlike Viagra ads in e-mail, a non-trivial population actually does want to join her network even if they understand it only exists for public relations purposes. The official Britney Spears' profile on MySpace has 135,385 links as of June 10, 2007. There are also numerous other profiles devoted to her; some are fan clubs, while others seem to be spam-like fronts for pornography sites using the popular and provocative singer's name and image as enticement.

The role of the SNS filter thus needs to be more than a marker of unambiguous spam. It needs to assist the user in assessing profiles by highlighting and clarifying the most salient features of an unknown contact, making it easier and more efficient for the human user to determine whether they wish to accept the invitation. This determination may vary greatly for the same contact across different users.

If all contacts were honest about their identity and intention, the user's task of determining with whom to connect would be much simpler. But they are not: the numerous spam senders mimic ordinary human users, creating credible profiles and even networks of reassuring, though fictional, friends. Detecting these deceptive entities is not impossible, but it takes considerable effort. Much of our work is focused on assisting this task – of highlighting the often subtle features that are characteristic of deceptive versus honest profiles and networks.

2. PROBLEM

Vaughan-Nichols noted that spam is almost impossible to define [21]. A penny stock ad is widely considered spam, but an extraneous advertisement from your bank might still be considered legitimate. Yet despite the gray area, spam has a clear enough definition that e-mail providers Google and Yahoo will try to filter it by starting from a master universal filter. Such master filters work for Google under the assumption unsolicited messages about medication, penny stocks, fake university degrees, and software discounts are universally undesirable. When Google misclassifies our mail, we correct it by toggling a binary spam flag. This approach towards e-mail spam is reasonable given a) we typically are not contacted by unknown ordinary people, and b) there is near universal consensus regarding which messages should be declared spam. But what happens when both of these assumptions fail?

In SNS, we can no longer assume that unsolicited likely means unwanted. SNS facilitate desirable unsolicited communications, opening a large gray area for "spam" classification. Should spam filters take on the role of sorting through the full gamut of desirable and undesirable unsolicited communications? We believe so.

We postulate that for SNS, the redefinition of spam filters should start by focusing on the sender rather than the message. Content analysis might be enough to discover a Viagra ad, but often in SNS we first need to judge character. Requests to join a member's social network are nearly content-less, containing only a link to the sender's profile. Thus, we are forced to first judge the sender by any available public information.

If we are still only detecting the presence of select categories such as penny stocks or pornographic webcams, we can simply redirect current content analysis techniques to the profile. But if we are rejecting a sender only because they are a celebrity, we are rejecting a *social prototype* [17] of a human rather than simply detecting one. Without the capability for machines to subjectively reason about people, we cannot adapt current spam techniques to SNS; Britney Spears and Viagra are similarly evaluated into the same binary world.

Filtering based upon wide category memberships is a far more complicated task than is possible using current spam filtering methodologies. As humans, we cast profiles into different types and categories, accepting those that appear interesting while rejecting those that do not meet our subjective criteria. Such assessment is prone to deception, as the malicious keep their intentions well hidden. Ideally, spam filters would help prevent deception and the uninteresting by assessing honesty and character.

Ultimately, a people-oriented AI reasoning engine is needed to interpret virtual identities and present the results. It would differ from current spam techniques in how people are evaluated (moving from templates to more holistic and necessarily subjective approaches) and how the results are presented (the machine's categories matching human categories). It could no longer interact with the user solely through binary spam flags; the ranges of unsolicited senders in SNS are far more graded than in e-mail. Desirability is dependent upon the recipient's current personal value system which in turn may change over time. Some might approve a link to Britney Spears simply because they think she's beautiful, while others appreciate her iconic status as a deviant. To account for the wide variety of preferences and criteria, we seek richer representations of people and content for the purposes of modeling social cognition and facilitating interaction with the results.

Creating an expressive socially minded representation of virtual personas for presentation to end-users and for evaluation by machines is non-trivial. As humans, we judge others on higher-level social rules than *is_penny_stock*. Disambiguating counter-culture from weird and undesirable can be difficult even for humans, let alone a machine.

Our long-term goal is to represent the character of senders using a notion of prototypes, or conceptual categories [13, 17], and more immediately using a selection of feature bundles. Prototypes are difficult to define as "a large proportion of our categories are not categories of things; they are categories of abstract entities. We categorize events, actions, emotions, spatial relationships, social relationships, and abstract entities of an enormous range: governments, illnesses, and entities in both scientific and folk theories, like electrons and colds" [17]. Thus, prototypes we might use in SNS are largely dependent on the goals of the user and the underlying system. If one chooses to reject a link to Britney Spears (see Appendix 1), is it because she is a celebrity (a high-level social characterization), or is it because she only unidirectionally broadcasts uniform information (a lower-level

[1] Britney Spears is currently a popular singer and cultural icon in the United States and worldwide.

characterization of network usage, or what we term a feature bundle)? Feature bundles are much closer to what we can extract without high-level reasoning and/or cognitive models. They are not the language commonly used by people in every day descriptions, but they are still useful the context of SNS while simultaneously easier to algorithmically generate. Examples of a feature bundles include "someone who sends more movie clips to their friends than they receive" and "someone with little public information available." They are likely to be the basis upon which we unconsciously generate our prototypes of the world, and therefore, if well-chosen, they can provide the necessary social cues to aid the user in navigating a virtual social sphere.

The situation is further complicated depending if one is solely trying to detect classical spam, or classify and separate ordinary humans into our own categories. In the case of classic spam, standard network-based metrics such as clustering coefficients [4] can work well without the need for prototypes or high-level characterization. However, the "spam" problem in SNS will soon be further complicated when many unwanted contact attempts come from real, ordinary social people rather than deceiving robots. In this scenario, forms of categorization have to separate out different classes of people that match the conceptual categories of the user's model of the world. Most users do not use clustering coefficients to assess who might be a desirable acquaintance. If a machine is able to accurately state that another member is central to the local punk rock scene, or that they share and pass on similar kinds of media, then we are approaching the end-user's mental model and reasoning methodology. Until our long-term goal of subjective machine assessment using relevant social prototypes is reached, our categories ought to be less subjective and more quantitative.

Some researchers have proposed that we can filter unwanted senders by injecting explicit or implicit trust values into the network [9, 11, 14, 19]. Such an act is a subtle blend between human detection and type characterization; we "trust" our friends' judgments to be valid across all dimensions. Assuming trust values provide a desirable statistic, these systems can only work reliably well in an open environment within the limited scope of friend-of-a-friend. As we compound multiple trust values to evaluate a node many hops away, our confidence in trust quickly diminishes as the nodes effectively become strangers [7]. This would not be such as much a problem in closed environments such as corporations, but in SNS, it is precisely these complete strangers that we desire to evaluate most.

Trust metrics are also problematic in that their definition is often one-dimensional. A single quantitative value cannot take into account how context and time changes the relationship between members. For example, we trust a friend to not intentionally send us a virus, but we may not trust them to not send us marketing information about their new company. With each hop they compound changing social practices and contexts in which the value was originally assigned, quickly abstracting their value into a higher dimensional space than their single dimensional can afford.

3. EXPERIMENT

3.1 User Characterization

We began our analysis by attempting automatic characterization of MySpace profiles using higher-level social categories. Specifically, our system describes a profile's valence in two independent dimensions: sociability and promotion. We evaluate sociability by the presence of information of social nature. A large number of personal comments, graphical customization, and other pieces of normally practiced social activity on MySpace yield a higher score. Promotion is evaluated by the amount of information meant to influence others, whether of political beliefs or of commercial nature. Typical e-mail spam would rate high in promotion, but low in sociability as they try to influence recipients without social dialogue. By contrast, a local rock band likely will score high in both dimensions since they actively communicate with their fan base on a personal level. Note we naïvely performed this experiment without a people-oriented reasoning engine, relying on basic machine learning techniques and a moderately small collection of feature bundles.

Further note that sociability does not simply refer to the amount of any information or content available. For example, we found that it is customary in MySpace to post a "thank you" message to another member's public bulletin board upon joining their social network. Surprisingly, this happens frequently even on profiles that have no intrinsic social value, i.e. pornographic webcams. We consider such messages to be somewhat sociable, but without additional personal content the cumulative social score for these generic messages would be very low. On the other hand, Britney Spears *could* score high despite being a commercial entity in the presence of personalized communications to and from her fan base.

We chose sociability and promotion because we believe the quadrants of their intersection could be the basis of initial expansion from spam/not spam to four useful categories:

Prototype 1: Low sociability and low promotion. This user might be a new member to the site, or might be a low-effort spammer who does not care about posing as something real. Without information to judge, we cannot tackle their classification.

Prototype 2: Low sociability and high promotion. This is typical of a promotional entity using SNS as a marketing opportunity. They only broadcast uniform information to their network, while simultaneously trying to expand its membership as much as possible. Examples include Britney Spears (who does not communicate individually with her network members), a Viagra ad, and a pornographic webcam.

Prototype 3: High sociability and low promotion. Such a rating is indicative of normal social-oriented humans. They connect and communicate with their social network on a personal level by posting pictures of themselves with their friends, results of random pop quizzes, and publicly host a suite of personal comments posted by their friends. Fortunately, they still constitute the majority of active SNS users.

Prototype 4: High sociability and high promotion. Unlike the generic marketing approach of Prototype 2, these promotional entities engage with their network on an individual basis. Often small-scale media producers (local bands, aspiring YouTube-based directors) fit this characterization by using SNS to connect with their audience.

To reason about virtual entities more holistically, we require some minimum of information to judge. Therefore, it is reasonable that we assumed network history to be a prerequisite for our analysis.

It is not without precedent; Boykin could well identify spam by analyzing the structure of social networks build from personal e-mail archives [4]. However, e-mail archives are private and incomplete. SNS already provide a rich source of public information in their network structure and collections of past actions. Any public profile (which we calculate to be 78.7% on MySpace as of December 2006) yields a user's entire social network, a self-made virtual identity, and a set of public messages sent to them by their verified network. The better and more complete the information, the more accurately we can judge the person. Fortunately for us, the current youth culture has fully adapted the desire to live public network lives [2,3].

3.2 Data Collection

We conducted an initial investigation to see if standard machine learning techniques could accurately predict the classification of MySpace profiles in sociability and promotion using a collection of features specific to the network and its culture. We tried to capture a spread of commercial, nonprofit, and noncommercial persona by picking MySpace profiles at random. MySpace exposes their user IDs as integers, which monotonically increase over time. This exposed database structure allows us to randomly choose profiles by algorithmically generating a list of pseudorandom numbers in the valid range of IDs, which we then fetch using an automated script. After collecting all the profiles, we hand rated each profile from one-to-five in sociability and promotion. A higher score in a given dimension corresponds to a higher valence. We will now refer each score by the variables *s* and *p,* representing sociability and promotion, respectively.

We only entered profiles into our dataset where at least $s>1$ or $p>1$ in order to process profiles with some minimum of information to judge. As we reached the thousandth profile, only 11% of our database had $p>1$; the majority were bands which already have a reliable special flag on MySpace. We know that the number of promotional profiles is increasing, but our data suggests MySpace still has far more social-oriented content than non-social (disregarding bands). Therefore, we focused on growing our promotional dataset specifically until we reached 400 profiles where $p>1$.

The 400 $p>1$ profiles were balanced against 400 profiles of $p=1$ for learning purposes. That is, we had 400 promotional-oriented versus 400 promotional-less. If we use the current real-world distribution, a random guess of $p=1$ would be correct 89% of the time. Given that we do not know if any of our features (to be explained) are meaningful, or if our dimensions are learnable, 90% accuracy is too close to a goal score. Therefore, we opted to balance the two sets by allocating 50% of the data to $p=1$. However, the 400 $p=1$ profiles were selected such that they maintained the same distribution in the sociability dimension as the larger data set contained where $p=1$. Table 1 shows the breakdown.

After obtaining the contents and rating of each profile, we further collected the profiles of each person's "top friends". Top friends are a special subset of friends that one can specifically select to be displayed on their main profile page. This is interpreted in the culture of MySpace as showing one's "best friends" [1]. We chose to use this subset for two reasons: 1) the full graph two hops from 800 profiles is especially large, and 2) we hypothesized that network statistics influenced by cultural practices will usefully highlight normal social processes.

Table 1: Distribution of profiles by sociability and promotional

	$s=1$	$s=2$	$s=3$	$s=4$	$s=5$
$p=1$	--	93	99	85	123
$p=2$	1	5	7	16	60
$p=3$	3	2	4	3	6
$p=4$	46	17	5	3	5
$p=5$	183	54	11	4	5

3.3 Feature Extraction

Determining the best features for our prototypes is non-trivial when their successful evaluation requires separating generic from personal relationships. We hypothesize that network-focused metrics are among the most important statistics to distinguish between usage types in SNS, for the purposes of both human detection and human categorization. The construction of individual profile is low cost compared to the potential of deception it brings in isolation. Yet, the construction of an entire network and its falsified chronology of activity is costly. Even when deception is not a problem, network activity (messages to and from friends) gives clues to a potential acquaintance as to what they might expect by joining the private network.

The structure of a network alone can be used for human detection. Boykin and Roychowdhury showed that e-mail spammers reliably have many edges but few wedges [4]. Wedges arise from shared communities and geography, a common feature which spammers lack. Yet a clustering coefficient alone cannot tell us that Tila Tequila [5], who tries to speak to as many strangers as possible, is not typical spam. In her case it would be better to search for different network-based cues and feature bundles, such as the bi-directionality in communication with her fan base, and whether her "friends" propagate her media to their friends.

We selected our features by thinking broadly about how people use MySpace. This includes information available on the user profile, as well as the comments written on one's top friends' profiles. Our choice of features reflects social trends on the site, such as the common use of easily detectable third-party content oriented towards MySpace profiles. Table 2 shows a hierarchy of our egocentric features, where "top *n*" refers a user's top friends, and "us" refers to the user in question. When we say "percent of our comments' hrefs that are unique," we are referring to hyperlinks found within our entire dataset to the same Internet address located in the comments posted by the user in question. Thus, it is possible many profiles in all of MySpace link to the same place, but we were unable to capture that in our subset of data. As a result, some of our features are inherently unreliable in our current configuration.

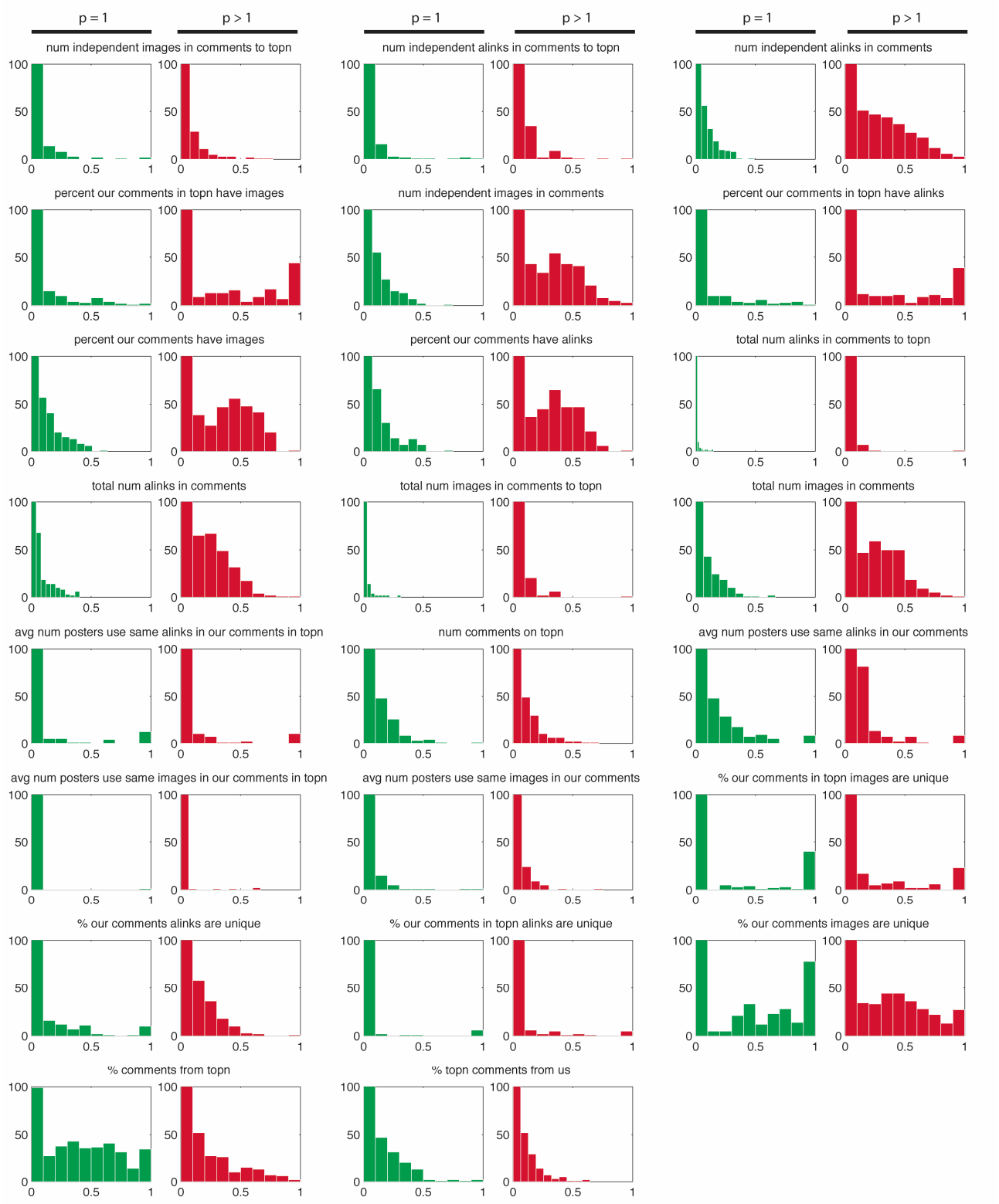

Figure 1. A histogram showing the results of network-based features. Each data point is put into one of two potential columns depending on its promotional score, independent its sociability score. This separates promotional entities (right and red) from the non-promotional (left and green). Note these specific distributions are a better shown by clipping the graph at 100 members on the y-axis as to visually concentrate the reader on the important details of the distribution while maintaining a small graph size for the purposes of publication.

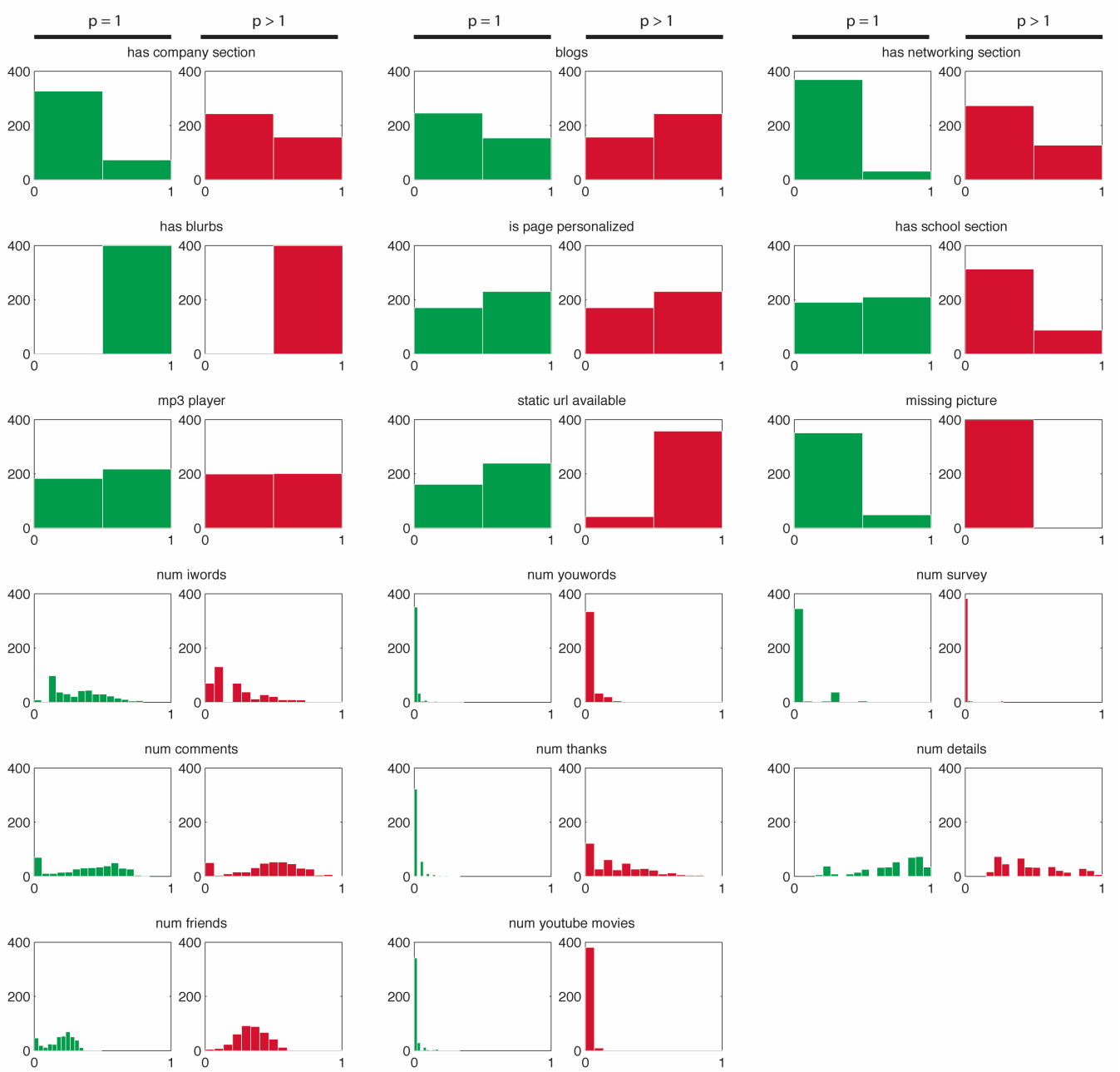

Figure 2. A histogram showing the results of profile-based features. Like Figure 1, each data point is put into one of two potential columns depending on its promotional score, separating promotional entities (right and red) from the non-promotional (left and green).

We normalized each feature from 0 to 1 so that all dimensions could be compared in the same linear space. Figures 1 and 2 show a histogram of the feature distributions of promotional-oriented profiles and those with no promotion. Despite a large bin around 0, most features display normal or power-law distributions. It is interesting that for several of the features, such as "percent our comments have images", the type of distribution changes depending if p=1 or p>1. Thus we have evidence promotional entities use the network differently than non-promotional entities. Our illustrated separation of data points is close to the notion of spam/non-spam in e-mail.

3.4 Machine Learning

As we did not know if our features or dimensions were learnable, we choose to survey many types of algorithms to see if any were suitable for our problem. We used linear regression, k nearest-neighbors, back-propagation neural networks (with varying number of hidden units and layers), and naïve Bayesian networks. Each algorithm ran multiple times with varying permutations of the following feature sets: profile-based, network-based, and mixed.

Table 2: Features extracted by category

Network/Comment Based

- percent of our comments that are from our top *n*
- percent of our top *n* comments that are from us
- percent of our comments' images that are unique
- percent of our comments' hrefs that are unique
- percent of our comments to our top *n* that have unique hrefs
- percent of our comments to our top *n* that have unique images
- average number of posters that use the same images in our comments to our top *n*
- average number of posters that use the same images in our comments
- average number of posters that use the same hrefs in our comments
- average number of posters that use the same hrefs in our comments to our top *n*
- total number of comments from anyone to our top *n*
- total number of images in comments
- total number of hrefs in comments
- total number of images in our comments to our top *n*
- total number of hrefs in our comments to our top *n*
- percent of our comments that have images
- percent of our comments that have hrefs
- percent of our comments in our top *n* that have hrefs
- percent of our comments in our top *n* that have images
- number of independent images in our comments
- number of independent hrefs in our comments
- number of independent images in our comments to our top *n*
- number of independent hrefs in our comments to our top *n*

User/Profile Based

- number of friends
- number of youtube movies
- number of details
- number of comments
- number of thanks
- number of survey
- number of 'I'
- number of 'you'
- missing picture
- mp3 player present
- static url to profile available
- has a school section
- has blurbs
- the page is personalized through CSS
- has a networking section
- has a company section
- has blog entries

Given 40 dimensions and only 800 data points (600 for training, 200 for testing), we feared the curse of dimensionality. We approximated feature selection using Principal Component Analysis (PCA) to reduce our space. We varied the number of dimensions kept with each learning algorithm, from 1 to 40.

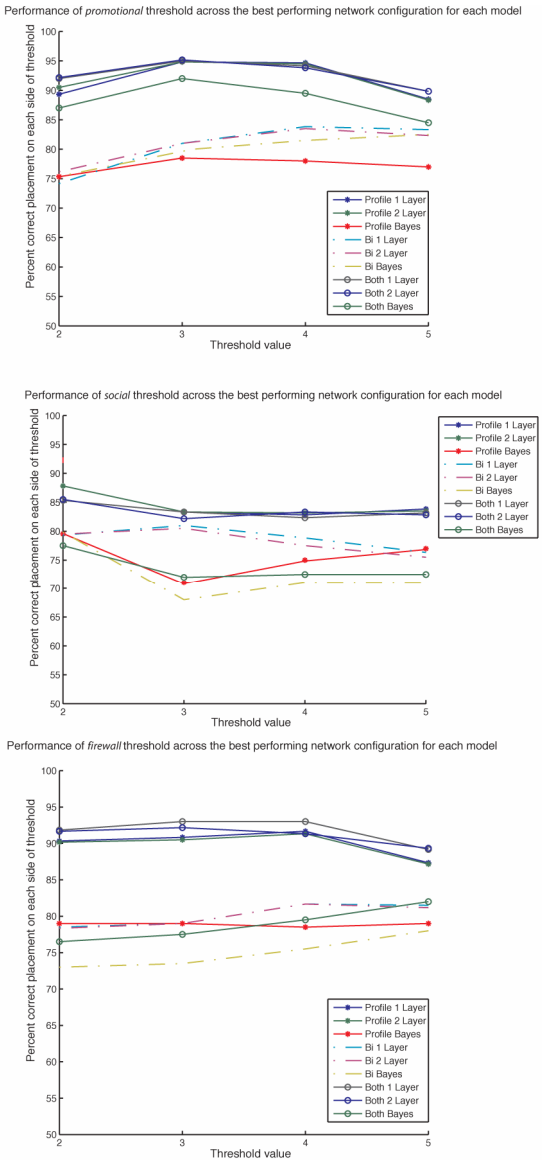

Figure 3. Graphs showing the best performance of each classifier permutation for each of the three threshold-based tests.

3.5 Results

In this paper we will only discuss the results of our neural networks and naïve Bayes experiments. Their scores were better or similar to our attempts with linear regression and *k* nearest-neighbors.

Our networks performed poorly in correctly classifying a profile in both dimensions simultaneously. The network did not do much better than 30-50% in any configuration, which is still better than random (see Table 3 for typical performance).

As we will later discuss, there was a large amount of subjectivity in the hand rating of the profiles. Due to time constraints, our hand rating only underwent a single pass per profile. Thus there is a high probability that another pass at the same profiles would result in the slightly different score, even from the same original

Table 3: Exact classification rates using two layer neural network and combined feature set on testing data

	s=1	*s*=2	*s*=3	*s*=4	*s*=5
p=1	--	0%	0%	0%	70.5%
p=2	43.5%	0%	0%	28.6%	7.7%
p=3	26.9%	0%	0%	0%	0%
p=4	--	--	--	--	--
p=5	55.6%	0%	0%	--	0%

reviewer. To handle this situation and get closer to how a human might expect to interact with a filtering agent, we created several new tests based upon a notion of thresholding. Our thresholding function seeks to correctly guess which side of given value (from two to five) a profile falls in a given dimension. For example, if our threshold is at three, and the data is actually one and we guess two, we would count that as correct because everything is on the same side of three. However, if we guessed three and the correct answer was two, our test would evaluate to false. The thresholding function reduces the subjectivity in our original ratings by fuzzing the scores while concentrating on a single dimension.

We created threshold tests to classify each dimension independently, in addition to a special "firewall" threshold that crosses both dimensions. Firewall is a special test that tries to represent the spirit behind current spam filtering, which is to block out non-social promotional entities. It is the same as the promotional threshold test except we also require $s>1$ (the profile is at least somewhat social). The user then sets the maximum promotional value a profile may score and still be let through.

All of our threshold tests unsurprisingly performed significantly better than the exact requirement tests, showing that at least something useful could be extracted from our features and dataset. For the firewall test, our performance ranged from 90-93%, with the best performance at t=4.

Surprisingly, we found that reducing dimensionality using PCA did not improve performance: much of the reduction actually gets performed by the trained network. This was also evident by the fact that fewer hidden nodes performed extremely well in our Neural Networks. Thus, we conclude the task may be inherently more linear or less multivariate than we previously assumed.

The best performance came from using both feature sets in a single layer neural network (Figure 3). However, this was only marginally better than using only profile-based features. We conclude that there is still value in including network usage statistics, but our profile-only features were good enough to get us most of the way there. The network-only tests fell between 78-83% accuracy, much lower than with the profile-based features. While this might seem discouraging when our goal is to use network-based features, we hypothesize that our preliminary feature set has much room for improvement by using more robust network statistics. For example, we did not include timestamps of comments in our features. The networks and comments of a "real" persona are built up organically over time, a process and resulting network and communication pattern that is difficult for spammers to mimic.

Profile scoped features will have a limited time that they can be considered useful in the spam/anti-spam arms race. We currently see a large increase in e-mail of image-based spam, simply because it is more costly for modern filters to handle. While spammer techniques will always adapt around the current detection technology, we believe a network-centric approach is ultimately more robust.

4. FUTURE DIRECTIONS

We believe we have identified a promising conceptual scaffold to filter solicitations in SNS by using the concepts of prototypes and feature bundles. Although our preliminary results are less substantial than we had expected, we believe the flaw to be in the choice of analytical techniques rather than the underlying network-centric approach. Our next step is to use more advanced techniques to analyze the network for the separate purposes of deception detection and human categorization.

As previously mentioned, Boykin and Roychowdhury have shown the clustering coefficient of a generated social network to be useful in fighting email spam [4]. They first examine the headers of an individual's email archive to approximate the actual social graph, then using its network properties classify users into white and black lists. While their methods could only be applied 47% of the time due to algorithmic constraints, when applicable it works fantastically well. Clustering coefficients are a promising example that network properties can at least usefully distinguish normal human behavior from the purely deceptive and malicious. Kimura et al. showed a similar technique can work well for search engine spam within trackback networks [15]. As we have already discussed, it remains an open question which network properties are appropriate given the changing subjective goals of a classification and the typical usage properties of a given site. Clustering coefficients are only useful if the culture of the network supports it.

We believe more research in passively generated statistics of SNS usage can get us much of the way there. Usage is influenced by pre-existing social conditions; we bring our cultural norms, communities, schools, geography, and friends into the networks we use. Sometimes local properties like geography can be a stronger force to grow the network than the network itself [20]. Some patterns, such as temporal rhythms [12], function well as markers of average human activity. More social science research into SNS is needed to distinguish the different types of users and cultures within a given network [1, 8, 12]. Such work is invaluable when algorithmically applied to detect humans and the various categories within them.

The features we choose directly impacts what we can predict and what we can show the user. If the desired categorization is too ambiguous or high-level, even the best classifier engine is likely to perform poorly. We chose sociability because we believed it matched the raison d'être of SNS; promotion reflected the growing misuse of SNS. We now realize they were inappropriate choices because their evaluation requires value judgments. For example, how sociable is (MySpace commercial entity) Britney Spears? Do personal responses from separate public relations interns constitute sociability? Do her "friends" need to actually know her in real life? As we further dive into the analysis of profiles, we uncover even deeper philosophical questions that challenge our assumptions and expectations of a virtual identity. Must only one mind to represent an entity? Does "it" need to be human? Does it need to be just one human, or can it be two humans and a dog?

What if it is clearly a human but is primarily about their business? Such questions highlight how arbitrary our current definitions might be as computer scientists when proposing generic anti-spam solutions.

Until we have reliable agents using a fine-tuned subjective cognitive model of the user, a better approach is to expose the end-user to a digestible form of the raw features and let them decide how to proceed. For example, the average ratio of messages sent to received might be enough for most people to filter a majority of profiles to their liking at a first approximation. This works because by itself it can be understood as meaningful social statistic: "Britney Spears" can be worthwhile as long as *it* usually converses back. Well-chosen fact-oriented metrics move value judgment to the only place they can be reliably interpreted across contexts: the user.

5. CONCLUSION

Computer-mediated communication is expanding our social boundaries. On the positive side, we in an information-based society are increasingly willing to make contact and develop relationships with people we have never met in person. On the negative side, we are also increasingly the recipients of spam and other unwelcome contact. The new social sphere of potential contacts, both desirable and not, is immense - and clearly we need technological assistance in sorting through it.

Our fundamental argument in this paper has been that this sorting needs to be more nuanced than the black and white, spam or not spam classification typical of most email analysis tools. We need to be able to classify a range of potential contacts to assist users of varying interests and tolerances in deciding which unknown contacts to accept and which to discard.

We attempted to do so by creating a model that could rate profiles in the dimensions of sociability and promotion. However, we quickly found that doing so requires placing a value judgment. When we, humans, were hand rating profiles to generate our data set, we often disagreed about what score a particular profile should take. For example, are political activists promotional, or is that only reserved for those selling something? If it is difficult for humans to agree on a particular rating due to subjectivity, how can we expect machines to perform the same tasks for us?

Only the user can decide if Britney Spears is spam. Yet the design of SNS and their associated services can speed this evaluation through digestion and presentation of information that would otherwise be hidden. Facebook has already begun the practice of publicly consolidating and aggregating activity of its users for consumption in its popular "News Feed" feature. However, it functions as a social radar at a literal level rather than a predictor of potential activity. If we expect the concept of social networking to become ubiquitous and of high utility, new interfaces will have to be built that highlight any past behavior indicative of future behavior. Without advanced AI, we presently advocate the presentation of facts without using subjective language or categorization.

We are confident that harbingers of promotional intent will emerge through analysis of network usage qualities. Regardless of our subjective follies, our histograms have shown at minimum that ordinary people and promotional entities have some differing character traits in network usage. At first this may not seem surprising, but the differing traits go beyond "how many people they attempt to befriend or contact." Clustering coefficients, gradients of bi-directionality in communication, and media sharing practices all give us insight into the behavior of entities that may be otherwise unreadable or too easily falsified. Future combinations of natural language processing with social network analysis have the potential to give an accurate prediction of what to expect from an unknown entity. It should be principally supported by examining an entity's role within the context of their friends and the culture across the entire site.

As John Keats famously wrote, "Beauty is truth, truth beauty, that is all Ye know on earth, and all ye need to know." In the vulnerable world of SNS, the truth may be ugly, but being able to reliably digest and present usage facts may be their only hope to preserve utility and curb chaos.

6. ACKNOWLEDGMENTS

We would to thank sponsors of the Media Lab, including News Corporation, for supporting this work. We also extend special thanks to Paulina Modlitba for helping perform the experiment and with the concept. Finally, we thank Greg Elliott, Nick Knouf, Tyler Maxwell, Christine Liu, Jonah Knobler, and Erik Sparks for their comments.

7. APPENDIX

[1] In SNS like MySpace, mainstream promotional entities are creating profiles and joining as many social networks as possible. Their connections are often used as a marketing opportunity to open a one-way communications channel without consideration for the recipients concerns.

8. REFERENCES

[1] boyd, d. Friends, Friendsters, and Top 8: Writing community into being on social network sites. *First Monday, vol 11, no. 12,* December 2006.

[2] boyd, d. Social Network Sites: Public, Private, or What? *Knowledge Tree 13*, May 2007. http://kt.flexiblelearning.net.au/tkt2007/?page_id=28

[3] boyd, d. Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life. *MacArthur Foundation on Digital Learning, Identity Volume (ed. David Buckingham). (in press)* MIT Press. 2007.

[4] Boykin, P. and Roychowdhury, V. Leveraging Social Networks to Fight Spam. *Computer, vol. 38, no. 4,* April 2005, 61-68.

[5] Caplan, J., Kingsbury, K., Jakes, S., Ressner, J., Rosenberg, G., and Walsh, B. Tila Tequila. *TIME Magazine*. December 16, 2006.

[6] Donath, J. (in review) Connecting the dots: Social networks sites as sketches of the future. *Journal of Computer-Mediated Communication*, 2007.

[7] Donath, J., and boyd, d. Public displays of connection. *BT Technology Journal. vol 22, no 4*, October 2004, 72-81.

[8] Fiore, A.T. *Romantic Regressions: An Analysis of Behavior in Online Dating Systems*. Master's Thesis, MIT Media Lab. September 2004.

[9] Gil, Yolanda, and Ratnakar. Trusting Information Sources One Citizen at a Time. *Proceedings of the First International Semantic Web Conference (ISWC).* Sardinia, Italy. June 2002.

[10] Girvan, M., and Newman, M.E.J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences USA. 99, no. 12.* 2002. 7821-7826

[11] Golbeck, J., and Hendler, J. Reputation Network Analysis for Email Filtering. *Proceedings of the First Conference on Email and Anti-Spam (CEAS),* 2004.

[12] Golder, S., Wilkinson, D., and Huberman, B. *Rhythms of social interaction: messaging within a massive online network.* http://arxiv.org/abs/cs.CY/0611137, 2006.

[13] Holland, D., and Skinner, D. Prestige and Intimacy: The Cultural Models Behind Americans' Talk about Gender Types. In *Cultural Models in Language and Thought,* D. H. a. N. Quinn (Ed.), New York: Cambridge University Press. 1987. 78-111.

[14] Kamvar, S., Schlosser, M.T., and Garcia-Molina, H. The EigenTrust Algorithm for Reputation Management in P2P Networks. *Proceedings of the 12th International World Wide Web Conference*. Budapest, Hungary. May 20-24 2003.

[15] Kimura, M., Saito, K., Kazama, K., and Sato, S. Detecting Search Engine Spam from a Trackback Network in Blogspace. *Lecture Notes in Computer Science, vol. 3684*, 2005.

[16] Krebs, V. *Data mining email to discover social networks and communities of practice*. Available from http://www.orgnet.com/email.html.

[17] Lakoff, G. *Women, fire, and dangerous things: What categories reveal about the mind*. University Of Chicago Press. 1987.

[18] Lenhart, A., and Madden, M. Social networking websites and teens: An overview. *Pew Internet and American Life Project*. 2002.

[19] Levin, Raph and Aiken. Attack resistant trust metrics for public-key certification. *Proceedings of the 7th USENIX Security Symposium*, San Antonio, Texas, January 1998.

[20] Marmaros, M., and Sacerdote, B. How do friendships form? *Quarterly Journal of Economics*. MIT Press. February 2006

[21] Vaughan-Nichols, S.J. Saving Private E-mail. *IEEE Spectrum. vol. 40, no. 8*, 2003. 40-44.