Organic or Organized? Exploring URL Sharing Behavior

Cheng Cao
Texas A&M University
College Station, TX, USA
chengcao@cse.tamu.edu

James Caverlee
Texas A&M University
College Station, TX, USA
caverlee@cse.tamu.edu

Kyumin Lee
Utah State University
Logan, UT, USA
kyumin.lee@usu.edu

Hancheng Ge
Texas A&M University
College Station, TX, USA
hge@cse.tamu.edu

Jinwook Chung
Utah State University
Logan, UT, USA
jinwook.chung@usu.edu

ABSTRACT

URL sharing has become one of the most popular activities on many online social media platforms. Shared URLs are an avenue to interesting news articles, memes, photos, as well as low-quality content like spam, promotional ads, and phishing sites. While some URL sharing is organic, other sharing is strategically organized with a common purpose (e.g., aggressively promoting a website). In this paper, we investigate the individual-based and group-based user behavior of URL sharing in social media toward uncovering these organic versus organized user groups. Concretely, we propose a four-phase approach to model, identify, characterize, and classify organic and organized groups who engage in URL sharing. The key motivating insights of this approach are (i) that patterns of individual-based behavioral signals embedded in URL posting activities can uncover groups whose members engage in similar behaviors; and (ii) that group-level behavioral signals can distinguish between organic and organized user groups. Through extensive experiments, we find that levels of organized behavior vary by URL type and that the proposed approach achieves good performance – an F-measure of 0.836 and Area Under the Curve of 0.921.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services

Keywords

URL; URL Sharing; User Behavior; Social Media

1. INTRODUCTION

URL sharing is one of the most popular avenues to share information on Twitter. Users can enrich their inherently limited length postings by inserting a URL pointing to an external resource such as a blog, video, or image. By doing so, many different viewpoints and additional context can be expressed through URL sharing. In the early days of Twitter in 2007, Java et al. already saw that about 13% of a collection of 1.3 million tweets included a URL [14]. Recent studies have confirmed the ongoing popularity of URL sharing on Twitter. In 2010, Boyd et al. found 22% of a sample of 720,000 tweets included a URL [3]. And in 2011, Rodrigues et al. found that nearly a quarter of 1.7 billion tweets contained URLs [21].

While some URL sharing is organic, other sharing is strategically organized with a common (perhaps, nefarious) purpose. And the differences between these two extremes – organic and organized – is often not a simple distinction. Consider the following examples:

- Figure 1 shows three users who tweeted the same URL `bit.ly/1dtous`, linking to a YouTube webpage related to the boy band One Direction. They all express their affections for the band in their tweets. It seems very likely they are the fans of One Direction, which explains that they spontaneously posted the same link. This *common interest* in a subject related to a link leads to the coincidence of multiple organically posted URLs.
- Figure 2 shows four more users who posted the same One Direction YouTube URL – `bit.ly/1dtous`. In this case, however, we can deduce that the users are participating in a voting campaign to attract the band to their hometown. The users are somewhat linked in this common desire.
- Continuing this theme, Figure 3 shows four additional users who have all tweeted a "vote" for Boston to attract One Direction. In this case, the voting behavior is suspicious: the tweets have highly similar text, and the latter three tweets were posted on the same day and the accounts names are quite similar. Were they organized to post the same URL? Are these accounts controlled by the same person? Is the first account "innocent"?
- Finally, Figure 4 highlights three users who engage in a clear example of a somewhat sophisticated organized URL spamming. Each user posts slightly different text and different appearing URLs, though ultimately all of the URLs redirect to the same destination URL – an advertising webpage. This coordinated behavior of URL sharing is fundamentally different from the first case of organic URL sharing.[1]

These observations motivate us to investigate URL sharing in social media. Our goal in this paper, *in the context of URL sharing*, is to automatically (i) identify *user groups* in terms of similar URL-sharing behaviors; and (ii) differentiate strategically *organized* and genuinely *organic* user groups, through the development of a URL-posting behavior based model. Our study aims to bridge the gap in the research of social media between user behavior, URL sharing, and general user similarity. The key insight motivating our

[1]These accounts have been alive for more than two years, which suggests the official Twitter spam policy has limited impact on this type of coordinated URL-sharing behavior.

CIKM'15, October 19–23, 2015, Melbourne, Australia.

DOI: http://dx.doi.org/10.1145/2806416.2806572.

Figure 1: One example of three users who have organically posted the same URL: bit.ly/1dtous

Figure 2: Four users seemingly post URL for a voting campaign

Figure 3: Four users suspiciously cooperate to post the same URL

work is that the publicly available URL-posting information can help model users with similar behaviors of URL sharing, and that group-level behavioral signals can characterize a group of users as organic or organized.

Studying similar user behaviors in URL sharing has important applications, on both bright and dark sides. On the one hand, based on the historical URL-posting information, the service provider (e.g., Twitter, in this paper) can effectively target groups of users and promote them their interested links. On the other hand, to purify and improve the information quality on their platform, it becomes imperative that the service provider can detect those organized behaviors of URL-sharing, such as campaign-like advertising, spamming, and other adversarial propagandas.

Given a group of organized users, on the one hand, we expect those users – no matter whether managed by a command-and-control structure or not – post URLs toward a common goal. We do not argue the goal has to be malicious, like the example in Figure 3. On the other hand, we focus on *group-level* behavioral evidences that can reflect their coordination. For instance, the users in Figure 2 were seemingly participating in the same "voting campaign", but they actually have different goals (the targeting cities) and their tweet content are quite distant. Even the users in Figure 3 seem suspicious, we need more evidences and should design a systematic framework for detection.

Concretely, we propose a four-step approach. We first formulate URL sharing based on its three key factors: user, URL, and the posting activity. Based upon such a model, we design a similarity measurement of user behaviors in URL sharing. Then, given the pairwise similarity function, we build a user graph model from which we identify user groups each of which contains users with similar URL-sharing behaviors. Next, on the group level, we characterize the organic and organized user groups based on the URL-posting behaviors of their members. Finally, we embed those characteristics into a classification framework to systematically distinguish organic and organized groups of users.

Figure 4: Three users coordinate to post the same advertising URL

2. RELATED WORK

Many recent studies have investigated URL sharing in social media, with different goals. One thread is about sharing intention, i.e., why people share links in social media. Suh et al. found that embedding URLs is one of the most important factors for increasing the *retweetability* of a tweet [24]. Smith et al. found that Twitter users add URLs to their tweets when discussing controversial topics, toward spreading information instead of conversing [22].

Another thread focuses on what people spread through URLs in social media, e.g., [20, 1, 16, 18]. These efforts have mainly focused on grouping similar messages or grouping users, such that the URLs provide additional context that may reflect the interests of the people posting these URLs.

Recently, several efforts have mentioned the dark side of URL sharing in social media. Stringhini et al. examined spam campaigns on Twitter that posted messages with URLs pointing to the same site [23]. Grier et al. defined a spam campaign as the set of Twitter accounts that spammed at least one blacklisted URL in common [13], and so Gao et al. did on Facebook [11]. Ghosh et al. studied the *link farming* on Twitter and found many participants are sybil accounts. Among those top link farmers, 79% have URLs linking to their external webpages [12]. More recently, Nikiforakis et al. explored the ecosystem of ad-based URL shortening services, and the vulnerabilities made possible by these services [17].

Since 2006, there have been many social network-based sybil defense methods proposed such as SybilGuard [27], SybilLimit [26], and SybilInfer [7]. Viswanath et al. pointed out most of those approaches are essentially *graph partitioning* algorithms [25]. They all made certain assumptions of the social network topology, used ground-truth information of trusted users, ranked all users, and determined who are sybils based on some cutoff. Rather than exclusively focus on spammers or sybils, our interest is to reveal groups of strategically organized users who engage in URL sharing with a purpose: some of the groups will post spam, but many others

spread propaganda, aggressively promote products or services, and generally engage in coordinated manipulation. Unlike spamming or subverting reputation, the users we consider can be genuine and legitimate, as in Figure 3. Additionally, our problem is more general in the sense that our approach can detect those "URL-posting based" sybils attacks or spam campaigns. We explicitly model and identify groups of users who have similar behaviors of URL sharing, and differentiate the organized and organic groups via a group-level classification framework.

3. INVESTIGATING URL SHARING

In this section, we propose our approach to explore similar user behaviors in URL sharing. In this context, our objective is to (i) formulate and collect user groups; and (ii) differentiate the organic and organized user groups. To tackle this problem, we formulate the concept of user group in the context of URL sharing, and focus exclusively on behavioral signals. We are interested in answering the following questions: How do we model user in URL sharing? How do we define and find user group? And how can we distinguish between organic and organized user group?

Toward answering these questions, our approach is structured in four steps. The main intuition is that the users from an organic group coincidently share similar interests on certain subjects so that they have similar behaviors of URL sharing. On the contrary, organized groups consist of users who plot to post URLs with a certain goal in common so that their behaviors conform to a notion of *co-operation* or *coordination*.

- First, we model the user in URL sharing, and design a similarity measurement to quantify similar behaviors between users.
- Second, we construct a user graph where nodes are users who have posted URLs and then identify user groups from the graph.
- Then, we extract group-level features to characterize organized user groups and organic user groups.
- Finally, we build a classifier for distinguishing organized and organic user groups.

We tackle each of these steps in turn, as follows.

3.1 Modeling URL Sharing

URL sharing on Twitter is fundamentally different from other popular activities such as *tweeting*, *re-tweeting*, and *following*. To systematically investigate users with similar behaviors of URL sharing, we model the three factors in URL sharing: the user, URL, and the action of posting. In the meanwhile, we need to pay attention on the following three issues. First, we should consider all URLs every user has ever posted on Twitter so that our model is general. Second, we should design a measurement that can quantify user similarity in URL sharing. Third, it should take the URL posting behavior into account. Such a measurement should be more specific towards URL sharing than other traditional user similarity based on profiles, social neighborhood, tweet content, and so forth. Therefore, we aim to model users engaged in URL sharing with two facets in mind: (i) Which URLs a user has ever posted; and (ii) How (s)he posted them.

3.1.1 User, URL, and Posting

Our key idea is that, in terms of the posting behaviors, a user can be characterized by all the URLs he has posted and how often he has posted them. For instance, a user who likes sports tends to share more URLs linking to sports websites than URLs of politics websites. A Japanese user is prone to post more URLs of Japanese websites than English websites. Two users who may have a realistic social relationship can have distinct tastes and preferences in URL sharing, but their posting behaviors (especially the posting frequency of different URLs) may reflect such a scenario. It is even more prominent for strategically organized users.

Formally, suppose we have a set of m users $\mathcal{U} = \{u_1, u_2, ..., u_m\}$. If u_i, in total, posted k different URLs $v_1, v_2, \ldots, v_k$, we define such an associated URL set of u_i as $\mathbf{p}_{u_i} = \{v_1, v_2, \ldots, v_k\}$. By aggregating all users, we get a URL set $\mathcal{V} = \{v_1, v_2, ..., v_n\} = \bigcup_{u_i \in \mathcal{U}} \mathbf{p}_{u_i}$. Therefore, each pair (u_i, v_j) can be seen as an action of posting a URL.

We introduce the function $f(u_i, v_j)$ that quantifies such a posting of v_j by u_i. Here, we give a straightforward definition $f(u_i, v_j) = PostingCount(u_i, v_j)$ where $PostingCount(u_i, v_j)$ represents the concrete *posting count* of every pair (u_i, v_j). Given a user $u_i \in \mathcal{U}$, we can represent u_i by an n-dimension vector $\mathbf{u_i} = (f(u_i, v_1), f(u_i, v_2), \ldots, f(u_i, v_n))$.

3.1.2 User Similarity in URL Sharing

We have two considerations to design a user similarity measure in URL sharing. On the one hand, if both users have posted overlapping URLs, the more intersections they have, the closer they are. On the other hand, we want to take the posting count into account. If two users posted the same URL many times, we reward the similarity between them. If not, we penalize them if they have quite distant posting counts for the same URL they have posted. Thus, we propose a measurement of user similarity defined as:

$$sim(u_i, u_j) = \sum_{k=1}^{n} \frac{ln(min(f(u_i, v_k), f(u_j, v_k)) + 1)}{|f(u_i, v_k) - f(u_j, v_k)| + 1} \quad (1)$$

We sum over all URLs to favor those pairs having posted many URLs in common. We pick the smaller posting count among two users ($min(f(u_i, v_k), f(u_j, v_k))$) as the pair-level scale of the posting count, and take the logarithm considering it can be quite large. We put the difference of posting counts as the denominator. We ensure a URL can contribute towards the similarity only if it was posted at least once by both users, yet it's different from traditional "cosine-like" measures as we explicitly emphasize the penalization.

3.2 Identifying User Groups

Given the definition of user similarity, the next question is how to find those users having similar behaviors of URL sharing. Given a set of users $\mathcal{U} = \{u_1, u_2, ..., u_m\}$, the task of *user group identification* is to find a collection of user groups $\mathcal{C} = \{\mathbf{c_0}, \mathbf{c_1}, \ldots, \mathbf{c_k}\}$ where $\forall \mathbf{c_i}, \mathbf{c_j} \in \mathcal{C}$, $\mathbf{c_i}, \mathbf{c_j} \subset \mathcal{U}$ and $\mathbf{c_i} \cap \mathbf{c_j} = \emptyset$.

The user similarity function can only locally measure the pairwise connection. Therefore, if we want to globally consider all possible users, adopting a graph structure is a natural choice. Extensive existing algorithms can partition a graph into *connected components*, which can fit our concept of user group here. In general, a *user graph* $G = (V, E)$ can be defined by $V = \mathcal{U}$ and $E = \{(u_i, u_j) | \forall u_i, u_j \in \mathcal{U}, weight(u_i, u_j) = sim(u_i, u_j)\}$. However, if we consider all possible user pairs, the resulting graph can be huge: it will have only one connected component where most connections are weak. Hence, we need a specific version of a user graph upon which we can extract our interested user groups.

3.2.1 The kNN User Graph

If we simply set a global threshold to filter out low-weight edges, we may lose important information. For example, the users from an organized group unnecessarily (and usually do not) post many URLs of popular websites — they have their own targets to spread. As a result, those users may be excluded from the graph due to the

lack of overlapping URLs with others, and finally we may obtain only a few big components in which people share popular websites.

Since we are interested in those organized users, our graph model should be able to grasp those "abnormal connections". Organized users do not often share various URLs, whereas the nature of coordination leads to their locally firm neighborhoods within the group. To retain such "conspiracies" as much as possible, we require a model that emphasizes the mutually steady neighborhood. Thus, we adapt the model of *k nearest neighbor (kNN)* graph.

The kNN user graph connects u_i and u_j only if u_i is among the k-nearest neighbors of u_j. Such a restriction retains only those k strongest neighborhoods, with less emphasis on the edge weight. A formal definition of the kNN user graph $G = (V, E)$ is $V = \mathcal{U}$ and $E = \{(u_i, u_j) \,|\, \forall u_i \in knn(u_j), \forall u_j \in knn(u_i)\}$. Here, $knn(u_i)$ is a function defined as $\{u_j | \forall u_j \in neib(u_i), sim(u_i, u_j) \in max_k(\{sim(u_i, u_j) | \forall u_j \in neib(u_i)\})\}$, where $neib(u_i)$ is the set of u_i's neighbors and $max_k(\mathcal{S})$ returns the k largest elements given $\mathcal{S}$. Now, since any node can have k neighbors, a group of users who post unusual URLs can still form a big component.

3.2.2 Extracting User Groups

It is a non-trivial task to extract a collection of user groups from the user graph. The concept of *connected component* in *graph theory* naturally matches our concept of user group, but we need two more considerations. First, we require that every group should have a compact size. Hence, we discard those small connected components (e.g., fewer than five members), and decompose those large components into smaller ones. Second, we hope the partition algorithm to be general, i.e., having been proved effective and efficient on many different types of graphs. Thus, we choose the well-known *Louvain method* [2], which is one of the most widely-used algorithms in *community detection* [8].

The Louvain method is a greedy algorithm to maximize the *modularity* of a graph structure. It starts by locally optimizing modularity for small communities, and then iteratively repeats to aggregate until reaching a maximum of modularity. Though the modularity optimization problem is known as NP-hard, the Louvain algorithm can run in $O(n \log n)$ in most practical cases so that it already has many mature implementations. Its maximum modularity ensures it conservatively segments big components only when it does contain multiple modularities.

3.3 Characterization: Organized vs. Organic

Given a collection of user components (groups) that we have extracted, our ultimate goal is to systematically distinguish organized and organic user groups. But before that, we need to characterize these two types of groups. Suppose we have a group of users who have similar behaviors of URL sharing, we want to feature to what degree they have organized behaviors when posting URLs.

First, we should ensure our features build upon the group level. Our biggest interest is the collective user behavior, which is fundamentally different from individual spamming or sybil behavior. Second, our features should target the behavioral signals of posting. We have seen in Figure 4 how easily the posters manipulated their tweet content to disguise themselves. Moreover, they unnecessarily follow each other as long as they post the same URLs. So, our insight here is those traditional features based on either text content or network structure become vulnerable for the organized user group in our context. On the contrary, what they cannot cover are the URLs they posted (they hope more exposures after all), the time-stamps they posted (they have to leave the records anyway), and their own profiles (they use public profiles so that others can see their postings). Therefore, in this section, we introduce nine group-level features that cover three posting-related aspects: *posted URLs*, *posting time*, and *poster profile*.

3.3.1 Posted URL-based Features

Our motivation is that, intuitively, the users in an organized group usually have a clear goal of promoting certain URLs. Thus, they have a relatively narrow selection of URLs they post, and each URL gets high-volume postings. Instead, an organic group usually posts a variety of URLs each of which has reasonable amounts of exposures. Therefore, we come up with two group-level features that capture the diversity of the posted URLs.

Suppose we have a user group $\mathbf{c} = \{u_1, u_2, \ldots, u_k\}$, and we know the set of URLs posted by every member, i.e., $\mathbf{p_{u_i}}$. Thus, we can extend it to the group level, as well as the posting count $f(u_i, v_j)$. Both can be formulated as the following:

$$\mathbf{p_{c_i}} = \cup_{u_j \in \mathbf{c_i}} \mathbf{p_{u_j}} \tag{2}$$

$$f(\mathbf{c_i}, v_j) = \sum_{u_i \in \mathbf{c_i}} f(u_i, v_j) \tag{3}$$

Based upon these definitions, we provide the following two posted URL-based features:

- Average Posting Count. We can calculate the *average posting count* per URL by the ratio of $|\mathbf{p_c}|$ and $\sum_{v_j \in \mathbf{p_c}} f(\mathbf{c}, v_j)$. By our motivation explained above, we expect organized groups have higher values of such feature than organic ones.
- URL Posting Entropy. Entropy is an important measure of *uncertainty* in *information theory*. Here, in our case, to describe the diversity of posted URLs in a group, we extend it to *URL posting entropy*, computed as:

$$H(\mathbf{c}) = -\sum_{v_j \in \mathbf{p_c}} \frac{f(\mathbf{c}, v_j)}{\sum_{v_j \in \mathbf{p_c}} f(c, v_j)} log \frac{f(\mathbf{c}, v_j)}{\sum_{v_j \in \mathbf{p_c}} f(c, v_j)} \tag{4}$$

  And based on the same idea, we suppose that organic groups have larger URL posting entropy than organized .ones.

3.3.2 Posting Time-based Features

One of the most important posting behavioral signals is the *posting timestamp* of every tweet. The poster has no access to tamper such information, making it a potentially robust feature. For each poster, we can collect a *posting time series* so we can compute all the *posting intervals*, defined as the temporal differences between consecutive posting timestamps. Our motivation here is: with the goal of promoting or advocating, the users from an organized group usually post tweets in a similar frequency, and the intervals tend to be short as they are eager to rapidly increase the exposures of their URLs. Therefore, we propose two group-level posting interval related features, measuring both the quantity and the deviation.

- Posting Interval Median. For every user, we can always quantify the posting interval by some temporal unit. Then we get the *individual posting interval median* by taking the median among all intervals. We use median rather than mean mainly because it is more robust to outliers. Moreover, since we care more about the group level, we take a further median over all members in a group. As mentioned, we expect organized groups have shorter interval medians than organic ones.
- Posting Interval Deviation. As said, our inference is that the organized accounts have similar posting manners, or even are bots manipulated by the same person. Thus, we can compute the group-level *posting interval deviation* given the individual interval median. Since an organic group is likely to post more randomly, we expect organic groups have larger deviations.

3.3.3 Poster Profile-based Features

Compared to an organized group, our main motivation here is that the users in an organic group have more various demographics. Given a group of organized accounts, if their goal is improper (e.g., advertising, spamming), they are mostly managed (like sybils) or hired (like for-pay turkers) by the same agent so they have close demographics. Even if their goal is relatively legitimate (e.g., propaganda, voting), their conspiracy attributes their enthusiastic interests on some common subjects, which can be reflected on their profiles to certain degree, too.

The profile information is one of the best publicly-available resources we can utilize to infer the diversity of demographics. We reify it into three aspects: total number of posted tweets, account registration date, and followers (friends) count. We calculate the group-level deviation of the total number of tweets, of the followers counts, and measure the average interval of the registration time.

- Tweets Counts Deviation. We count the total number of tweets an account has posted ever since the beginning, and take the deviation among all accounts in a group. The larger deviation means the more variety, so we expect an organic group has a larger *tweets counts deviation* than an organized one.
- Followers Counts Deviation. We record the count of followers for every user in a group. The count can be dynamic over time so we take the median. Then we take the deviation among all members. We believe the count of followers is more reliable than of friends simply because it is more difficult to fabricate. Again, we expect organized groups have lower deviations.
- Registration Interval Median. It is similar to how we computed the *posting interval median*. We look into the registration date of each user so that we have a *registration time series* for every group, and then we take the median. Our idea is that the registration interval is one of the most direct evidences for those fabricated or hired accounts. We prefer median to mean because the latter one is too sensitive to outliers. We expect organized groups have smaller *registration interval medians* than organic groups, similar to the reason for the posting interval median.

Then, we come up with two more features related to the registration date. We define the *user lifespan* by counting the temporal span, in terms of days, from the day an account registered to his latest posting date in our dataset. The motivation here is that the user lifespan is one of the most important user profile information and it should be quite random for users from an organic group. Instead, the organized users are created for certain URL-promotion mission, so they all tend to have short lifespans. Another idea is that Twitter may already have detected the suspicious behaviors from organized users and suspended them, leading to short lifespans too. Hence, we provide two features to characterize the group-level user lifespan via the quantity and the deviation.

- Poster Lifespan Median. We take the median of all accounts' lifespans in a group. As said, we suppose an organized group generally has a shorter *poster lifespan median*.
- Poster Lifespan Deviation. Every member from a group has a lifespan so we can calculate the group-level deviation. We expect organic ones have larger *lifespan deviations*.

3.4 Classification: Organized vs. Organic

Recall that our ultimate goal is to automatically discern organized and organic user groups, in the context of URL sharing. Given the features in the previous section, it becomes quite natural that we embed them into a classification framework. To choose appropriate classification algorithms, we should guarantee: (i) the algorithm has been widely used and maturely implemented; and (ii) we need to test on multiple algorithms.

We choose 4 well-known classification algorithms in our framework: *Random Forest* [4], *Naive-Bayes Decision Tree* [15], *Sequential Minimal Optimization* [19], and *Additive Logistic Regression* [10]. We notate them *RandomForest*, *NBTree*, *SMO*, and *LogitBoost*, respectively. NBTree is a *decision tree learning* algorithm in which the tree leaves are naive Bayes classifiers. SMO algorithm is used for training the support vector classifier and has been implemented in many existing SVM libraries like LIBSVM [6]. LogitBoost can be seen as a variant of AdaBoost [9] that adapts *logistic regression* techniques. All of them have mature implemented packages. With different theoretical foundations, these four candidates can well-serve the testing algorithms in our experiment.

4. EXPERIMENTS

We present our experimental studies in this section. We first introduce our data. Then, we describe how we identified all the user groups we formulated. Third, we provide details how we collected the ground truth. Finally, we show our analysis results towards distinguishing organized versus organic user groups.

4.1 Data

We deployed a tweet crawler via the official Twitter Streaming API from October 2011 to October 2013. Since our main interest is URL sharing, we only collect tweets that contain URLs. The API provides a 1% sample of all published tweets, but our 24-month uninterrupted crawling gives us 1.6 billion *"raw URLs"* posted by 136 million accounts. Raw URLs are those URLs in their original format when posted, without any further processing (e.g., resolution) after being crawled. Due to either irregular typing or the URL shortening service, many raw URLs become inaccessible or actually link to the same webpage. For instance, we have both `WWw.TwiTtEr.com` and `tWITTER.com/` in our dataset, as well as `bit.ly/1a8jUOr` and `bitly.com` both of which direct to the same destination.

To address such an issue of *URL variants*, we need to resolve for all URLs. Since resolving billions of URLs can be expensive, we focus on URLs that appeared at least 50 times. We resolve through standard HTTP requests and record the landing long URLs. The summary of our dataset is shown in Table 1: 47 million accounts have shared 1.6 million raw URLs within 445 million tweets. 82% raw URLs get resolved to nearly 1.2 million distinct long URLs. 869 thousand accounts have generated at least 50 postings.

Given a resolved long URL, we ignore all its post parameters as a reasonable approximation to the URL. We call the remaining part *URL domain name* and we obtain 166,000 unique ones. Compared to a complete URL, we believe the domain name has better interpretability because it can conceptually represent a "website". For users, we exclude those who just occasionally share URLs, i.e., less than 50 postings. To model the user, we decide to also use URL domain name as the dimension. One reason is that using original long URL may result in extremely sparse user vector given enormous dimensions. Another reason is a user group can be better interpreted if each member corresponds to some "website" instead of a long HTML page link.

4.2 Collecting User Groups

To construct the kNN user graph, a non-trivial issue is how to choose an appropriate k. We adopt the idea in [5] to pick k roughly equal to $ln(|\mathcal{U}|)$. Thus, 869,571 users give us k of 14. In the end, we obtain a user graph containing 216,523 nodes and 3,862,116 edges. This graph contains 12,251 connected components most of

Table 1: Dataset Summary

# Raw URLs (Resolved)	# Tweets	# Unique long URLs	# Unique domain names	# Posters (Having at least 50 postings)
1,617,234 (1,327,729)	445 million	1,199,930	166,107	47,658,839 (869,571)

which are small: only 2,150 components have no less than 5 nodes and just 36 components are bigger than 100. We filter out those tiny components smaller than 5, and exhaustively decompose (if possible) large ones bigger than 100 to ensure maximal modularity. Eventually, we identify 2,775 groups, together including 192,719 users. Among those 2,775 groups, we find around 40% groups are smaller than 10, and nearly 90% groups are bounded by 100 users. The largest one has 14,080 users.

How can we find a way to see whether our identified user groups are "meaningful"? And how to interpret and measure it here? Naturally, the most direct evidence is our groups maintain closer manners of URL sharing behaviors than "a random user group", yet we need a way to measure it. The entropy of posted URLs in a group is one of the most typical properties that capture the similar URL sharing behavior, as explained in Section 3.3. Thus, first, we simulate a collection of user groups with the exact same sizes via randomly picking users from our dataset. Then, we compare the distributions of the URL posting entropy for our collected groups and the simulated ones. The result is in Figure 5, and we clearly see the difference. Around 20% of our groups have a zero entropy, i.e., all the members posted the same one URL all the time, and the median is about 1.5. On the contrary, none of the simulated groups have a 0 entropy, and the median is around 4 while 80% of our groups are below it.

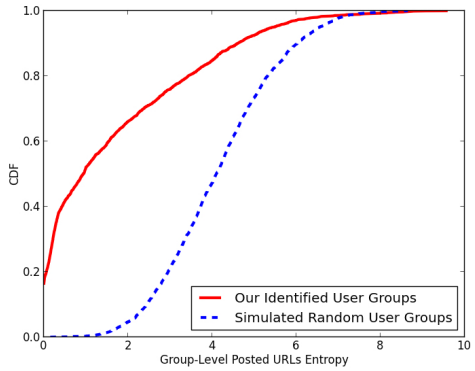

Figure 5: The URL posting entropy CDF of our collected user groups, compared with a collection of simulated random groups

Besides the natural evidence from those URL-posting related features, we hope to see more different evidences showing our collected groups are meaningful. Here, we adopt the group-level *language usage entropy* to measure certain "homophily" among all members in a group. The intuition is that if a set of users have similar selections of URLs (embedding in their tweets), then their language choice in their tweets should be similar. Conversely, two users who have distinct language backgrounds would hardly overlap many URLs. Therefore, for each group, we aggregate all the published tweets (with or without URLs, having valid language usage information), count the usage frequency, and calculate the entropy. To compare, again we simulate the user groups with the exactly same sizes. We compare their distributions in Figure 6, and find the contrast is apparent.

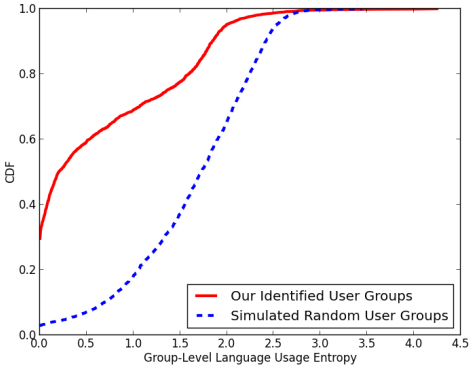

Figure 6: The language usage entropy CDF of our collected user groups, compared with a collection of simulated random groups

About 30% groups in our collection whose language entropy is 0, i.e., all members always use the same language writing tweets. 90% of our groups have entropies less than 2.0. In contrast, in the simulated collection, almost none have 0 entropy, and the median is between 1.5 to 2.0 where 80% of our groups are below it. This comparison shows the users in our identified groups have much similar language usages. Recall that our method of computing user groups has not used any user information on language usage, which demonstrates the potential of extending our approach into general problems related to user similarity.

4.3 Ground Truth

To test our proposed features in Section 3.3 on characterizing organized and organic groups, we need a set of groups with known labels of either organized or organic. Since there is no such existing ground truth, we randomly pick 1,000 of our identified groups, and manually check each of them as follows.

4.3.1 Manual Labeling Setup

Our labeling for each group is a two-tier task: (i) categorizing the content of the URLs the members posted and inferring the purpose that they posted URLs; (2) and rating to what degree we think their behaviors were organized or organic.

The rating task directly connects to our interest here, but the categorization task can help us interpreting the similar user behaviors. Our focus is to collectively seek for group-level evidence that reflects the coordinated behaviors of posting URLs, not just individually inspecting each account. In particular, we examine the accounts in each group through looking into their: (1) tweets (e.g., all the posted URLs, landing webpages, and tweets that contain URLs); (2) account profile (e.g., self-introduction, name, avatar, language, geo location); and (3) posting pattern (e.g., the URL shortening pattern, textual pattern in tweets, posting timeline pattern). If a group is too large, we randomly pick at least 5 users with accessible information to judge. Judges make decisions without knowledge of the proposed features in Section 3.3.

To measure to what extent a group is coordinating can be subtle. Our ratings will be the scores between 1 to 5, the larger the more

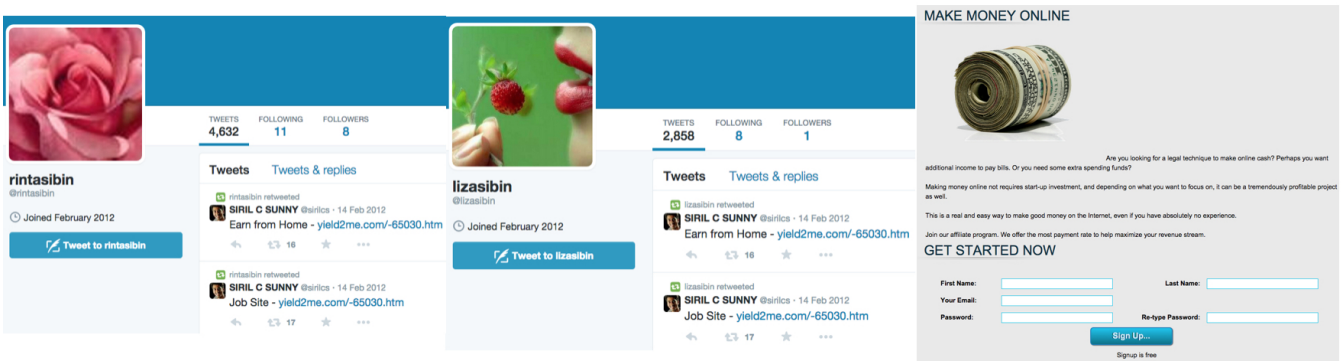

Figure 7: Example users whose group we categorize into "spam" and think is organized

Table 2: The distribution (percentage) of twelve categories that we have labeled for our user groups

advertising	spamming	app-auto-generated	entertainment	social-media	adult	news	blog	follow-back	public-information	propaganda	voting
318 (41%)	195 (25%)	171 (22%)	66 (8%)	33 (4%)	22 (3%)	20 (3%)	10 (1%)	8 (1%)	6 (0.8%)	6 (0.8%)	4 (0.5%)

suspicious towards an organized group. Then, we transform our scores to the label of either organized or organic. It is an organized group if its score is above 3, and organic if below 3. If its score is 3, we inspect it again. By such a 5-scale rating, we avoid curtly labeling a group is organized or organic.

We assign 3 annotators the exactly same set of 1,000 user groups, and separately ask them to manually label each group as organized or organic. Since the number of our annotators is odd, if the individual decisions of our 3 annotators have a clear "major voting", we take it as the final result. The rating usually has an obvious favor (below or above 3), and we accept a category if it has been mentioned by at least two annotators. Otherwise, a fourth annotator adds a label and a rating. We finalize a category or rating only when it has two or more endorsements, and count a user group has valid labels only if it finally obtains at least one category or rating.

4.3.2 *Categorizing a User Group*

As said, our categorization aims to infer the common intent of a group of users who have similar URL-sharing behaviors. The idea is that organized groups have a much clearer goal in common than organic groups, even though the purpose can be either improper (e.g., advertising, spamming) or legitimate (e.g., self-promoting, preaching). In practice, we do find sometimes speculating the intention is not obvious, so we change to summarize the content of posted URLs and tweets in this case. In the meanwhile, we find usually categorizing a group is more difficult than rating it, especially if the members post quite various subjects or many non-sense words. Finally, among all the 1,000 randomly picked groups we have judged, 773 ones have been labeled by at least one of the 12 categories in Table 2.

The category of *advertising* emphasizes the intention of posting URLs. It includes everything about absorbing viewer's attention on the URL, such as marketing, funds-raising, or even self-promotion of personal websites or uploads. We have seen instances that all members in a group shared URLs linking to the exactly same news articles or blogs. We still think such a group is suspicious to advertising. Moreover, the irrelevance between the content of tweets and the linked webpages can also indicate the poster is advertising, and one example is shown in Figure 8.

We make the assumption that Twitter's official suspension indicates *spamming* activities to some extent. In addition, if the browser or the shortening service warns when we click into a URL, we believe it is a spam URL. Other than that, we label spamming if the URL links to some typical spam webpage of "phishing" or "malware". For instance, we categorize a group full of users like the left-hand two in Figure 7 into spamming, because their posted URLs point to the phishing website shown in the rightmost subfigure.

Figure 8: Example users whose group is organized

App-auto-generated is a special category. All users in such a group posted tweets (containing URLs) automatically generated by some external app or service. Usually it happens when the poster is unconscious or at least not intentionally doing so, e.g., using own Twitter account to log in some mobile game. For *entertainment*, we find groups of fans who posted URLs about their supported artists like we have seen in Figure 1, 2, 3. Sometimes they can be *voting* (like in Figure 2 and 3), but many are just bonded by fan's passion. Other entertainment groups often talked about music, video game or anime. *News* are mostly about politics and technology. *Propaganda* usually relates to politics and religion.

Social-media and *blog* stress the source website. These websites can host various webpages, and typical examples are `youtube.com`, `instagram.com`, and `blogspot.com`, etc. This also explained we found many such groups. The groups of *public-information* posted links of public information like traffic or weather. The posters usually are accounts managed by regional institutions.

Among all these 12 categories, some are capturing the purpose of posting (e.g., advertising, spamming, propaganda), and some are summarizing the content of posting (e.g., entertainment, social-media, news). A user group can have more than one category, e.g., mixed with spam, adult, and advertising. In total, from Table 2 we can see advertising and spamming together take up 66% appearances of all categories. This observation reflects our idea that the

kNN graph model, emphasizing the locally pairwise connection, can potentially retain many abnormal behaviors.

Another interesting finding is the category of app-auto-generated surprisingly occupied 22%. Its substantive existence leads to a new thread for future work: the *user-unconscious* posting behavior in social media, caused by some external application. What security issues could it raise? Do most posts contain URLs? If yes, can our study of URL sharing be an entry point?

4.3.3 Rating a User Group

For rating, what has been posted matters less, and we care more on who have posted and how. As mentioned, we look over the profiles of all accounts from the same group. We seek for similar patterns between them in self-introduction, name, avatar picture, etc. We rate above 3 if we see highly overlapped textual patterns in their tweets such as the posting date, hashtag, language and URL shortening. Figure 8 includes two accounts with identical posting timeline, tweets content, and similar names. Their group is the instance that we rate 5. It becomes even more dubious when the members always retweet each other's tweets. A typical "rate-5" example is in Figure 7. The users, who have extremely close posting timelines and names, always retweeted from the other account.

Our rating focuses on the collective behavioral evidences of coordination, independent with the posting purpose. For example, we find most groups of entertainment are users spontaneously talk about the subjects they like, but it becomes suspicious when the URLs link to the same music video page on Youtube. Most groups of public-information have fairly legitimate goal of posting, but naturally many accounts are centrally managed so they still have many common patterns.

If all members in a group have been suspended (and we label it as a spamming group), we cannot access their accounts so we conservatively leave them unrated. For those groups of app-auto-generated or social-media, we mainly judge their postings are whether advertising-oriented or unconscious, and we usually find most of them fall into the latter scenarios. The groups of news, social-media or blog often share popular news, blog article, and online forum. We rate them low except we find they advertised their own-related (person or institution) webpages.

4.4 Experiments: Organized vs. Organic

4.4.1 Analyzing Our Labeling

As introduced, we finalize the labeling result for each user group via adopting the major vote over our 3 annotators. We find 986 groups that have been labeled at least one of the categorization and rating results, and 602 ones that received both. The fourth annotator agreed on 871 (88.3%) and 520 (86.4%) ones, respectively. Finally, we obtain 815 groups with ratings where 325 (40%) are organized and 490 (60%) are organic.

Those 602 groups with both information of category and rating give us the opportunity to understand the following questions: Which subjects our collected groups mostly talked about? What kinds of content the organized groups and organic groups usually posted? Are they different? Can we see a relation between the levels of organized behavior and the types of URL?

We aggregate the average rating each group received by the group category, and plot the CDFs of our ratings for all 12 categories in Figure 9. We can see the distributions take on clear gradients. If we look at the median, 12 categories are evenly divided by the rate of 3. The groups of spamming, adult, and follow-back are most likely as organized, followed by advertising, public-information, and propaganda whose distributions are quite close. Then, blog, voting, entertainment and news have similar distributions, most of whose groups are organic. Social-media and app-auto-generated groups are the least likely organized (recall how we rate them). These observations reveal the connection between the level of organized behavior and the inherent content of different categories. For spamming, adult, follow-back, advertising, they essentially include those improper ingredients that motivate more organized behaviors. On the contrary, the category of app-auto-generated is user-unconscious, and blog, news, entertainment and social-media naturally contain more legitimate activities, so their groups tend to be more organic.

4.4.2 Classifying Organized and Organic Groups

We are ready to build the classifier. In terms of posting behaviors in URL sharing, can our proposed group-level features distinguish the organized and organic groups? Which features work best? Which classification algorithm performs best?

To address the issue of class imbalance, we give a hybrid solution that combines undersampling of the majority class and oversampling the rare one. In the end, we have 406 organic and 406 organized groups, and we normalize the values of each feature to the interval of 0 to 1. To evaluate, we do 10-fold cross validation and focus on precision, recall, F-measure, and ROC area. We first consider the two classes are equally important and take the average. The results are in Figure 10.

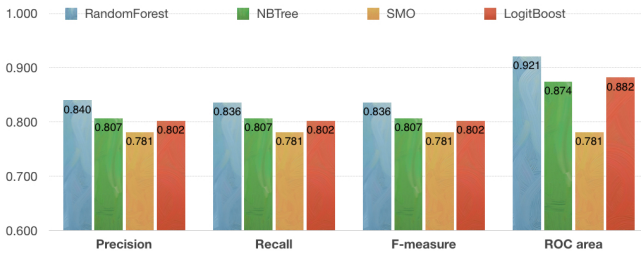

Figure 10: Evaluation results by four classification methods

The performance achieves around 0.8 no matter which method or measure we choose. RandomForest works best with high F-measure (0.836) and ROC area (0.921). All the results suggest the potential of our approach for distinguishing organized and organic user groups, purely based on behavioral signals in URL sharing.

In reality, detecting organized user groups becomes more important. We especially want to find out organized groups as many as possible, corresponding to the measure of recall. Thus, we further show the recall result for the class of organized in Figure 11. We see those two decision tree based algorithms outperform the other two, and even better than their own averaged results in Figure 10. This observation hints us we can prioritize decision tree based algorithms when we solve such a problem in practice.

Another investigation is the feature impact. We select two popular measures — *Chi-squared* and *Information gain* — to evaluate each feature with respect to the class. The full ranking is in Table 3. We see the rankings by chi-square and information gain are almost identical. We have 3 interesting discoveries. First, the two URLs-based features always rank the top 3. This suggests the URL-related property is the most reliable aspect when we study URL sharing. Second, if we have two features from the same type (e.g., lifespan deviation and median), the deviation feature always has more influence than the median one. One possible explanation is that the median value is too sensitive compared to the dispersion value. This tells us those features derived from a relative measure can be more robust than those from an absolute measure. Third,

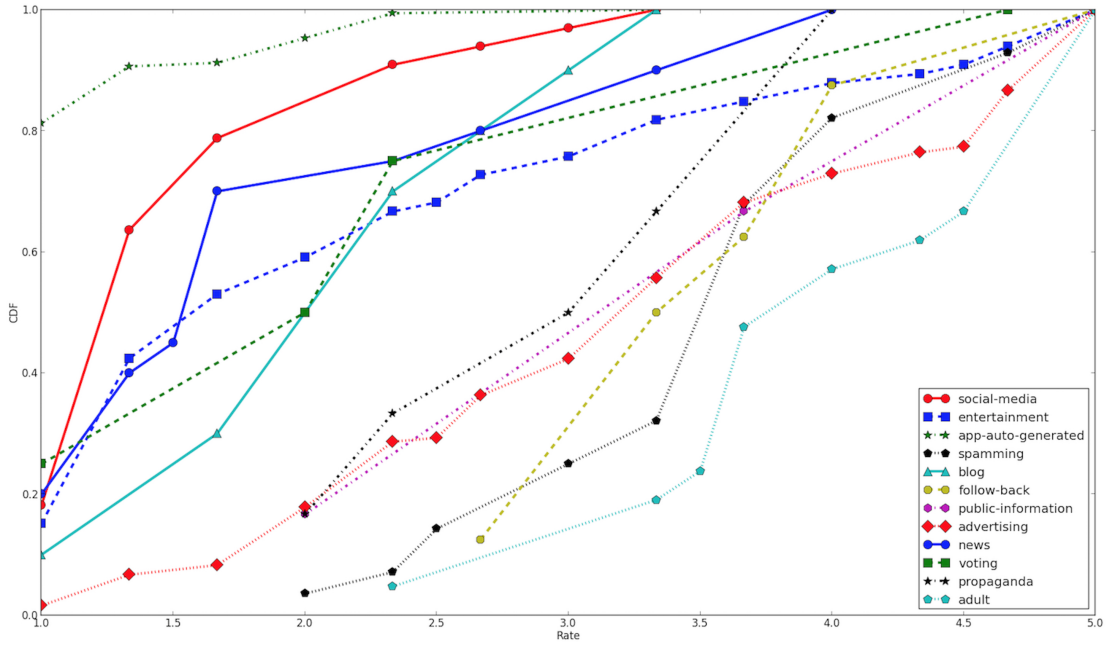

Figure 9: The CDFs of our ratings for all 12 types of group categories

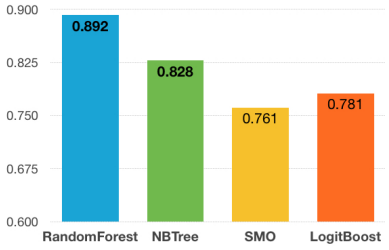

Figure 11: The Recall results for the class of organized group by four classification methods

the feature of poster lifespan deviation performs the best. It favors our intuition in Section 3.3: User lifespan is one of the strongest signals to capture the diversity of poster demographics.

Table 3: The rankings of the feature impact measured by Chi-Squared and Info Gain

Chi-Squared	Info Gain
Poster Lifespan Deviation	Poster Lifespan Deviation
Average Posting Count	Average Posting Count
URL Posting Entropy	URL Posting Entropy
Registration Interval Median	Registration Interval Median
Poster Lifespan Median	Tweets Counts Deviation
Tweets Counts Deviation	Poster Lifespan Median
Posting Interval Deviation	Posting Interval Deviation
Posting Interval Median	Posting Interval Median
Followers Counts Deviation	Followers Counts Deviation

We conduct two more informative comparisons of organic versus organized groups. One is about the number of distinct URL domains that a group has posted. Our idea is that organized groups tend to have a much tighter selection of URLs to post than organic groups, mainly due to their specific common goal of posting. In Figure 12, we find the significant gap between two classes. More than 20% of organic groups have mentioned at least 100 different URL domains in their tweets, and yet the median for organized groups is merely no more then 10. 20% of organized groups have posted only one kind of URL domain all along.

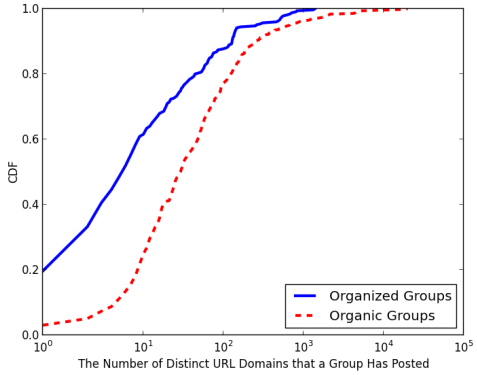

Figure 12: Organized vs. Organic: # URL domain names

The final exploration comes back to Twitter itself. We would like to see how the Twitter's official monitoring reacts for the two types of similar user behaviors in URL sharing. So, we calculate the *group-level suspension percentage* in each of our groups, i.e., how many accounts in a group have been suspended. Recall that we have excluded those groups whose members were all suspended. We have two interesting observations from Figure 13. On the one hand, the distributions of our two types of groups are quite discerning. We find 80% of organic groups have less than 10% suspension percentage, where 80% of organized groups are more than it in contrast. Twitter's suspension is for individuals, yet it still reflects on our user groups formulated through URL sharing. On the other hand, we find the official suspension still has limited impact on the organized posting behavior: the median is just around 30%. These

findings suggest the complementary potential of our investigations on organized user behavior in URL sharing.

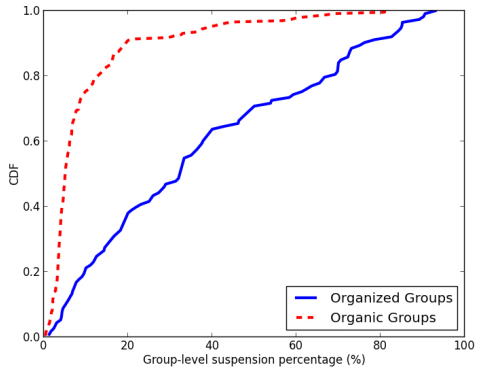

Figure 13: Organized vs. Organic: group-level suspension ratio

5. CONCLUSION

In this paper, we are interested in exploring users with similar behaviors when they share URLs on Twitter. While some users organically share common interests on certain websites, some are organized to aggressively promote the same URLs towards a common goal. This motivates us to tackle the problem of distinguishing organized and organic users in the context of URL sharing. Focusing on the behavioral signals of URL-posting, we propose a four-step approach to model, identify, characterize, and classify those two types of user groups. We test our approach on four different classification algorithms and in most cases it performs good in terms of precision, recall, F-measure, and ROC area. Random Forest algorithm works best with 0.921 ROC. Our experimental analysis demonstrated the capability of our approach for (i) understanding users with similar URL-sharing behaviors; and (ii) distinguishing the level of organized user behaviors in URL sharing.

6. ACKNOWLEDGMENT

This work was supported in part by AFOSR Grant FA9550-12-1-0363. Any opinions, findings and conclusions expressed in this material are the author(s) and do not necessarily reflect those of the sponsors.

7. REFERENCES

[1] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *WWW*, 2012.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[3] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS*, 2010.

[4] L. Breiman. Random forests. *Machine learning*, 45(1), 2001.

[5] M. Brito, E. Chavez, A. Quiroz, and J. Yukich. Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. *Statistics & Probability Letters*, 35(1):33–42, 1997.

[6] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *TIST*, 2(3), 2011.

[7] G. Danezis and P. Mittal. Sybilinfer: Detecting sybil nodes using social networks. In *NDSS*, 2009.

[8] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3), 2010.

[9] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *ICML*, 1996.

[10] J. Friedman, T. Hastie, R. Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2), 2000.

[11] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and characterizing social spam campaigns. In *SIGCOMM*, 2010.

[12] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi. Understanding and combating link farming in the twitter social network. In *WWW*, 2012.

[13] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @ spam: the underground on 140 characters or less. In *CCS*, 2010.

[14] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Joint 9th WebKDD and 1st SNA-KDD Workshop*, 2007.

[15] R. Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, 1996.

[16] J. Lehmann, C. Castillo, M. Lalmas, and E. Zuckerman. Transient news crowds in social media. In *ICWSM*, 2013.

[17] N. Nikiforakis, F. Maggi, G. Stringhini, M. Z. Rafique, W. Joosen, C. Kruegel, F. Piessens, G. Vigna, and S. Zanero. Stranger danger: exploring the ecosystem of ad-based url shortening services. In *WWW*, 2014.

[18] J. Park, M. Cha, H. Kim, and J. Jeong. Managing bad news in social media: A case study on domino's pizza crisis. In *ICWSM*, 2012.

[19] J. C. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods*. MIT press, 1999.

[20] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In *ICWSM*, 2011.

[21] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and V. Almeida. On word-of-mouth based discovery of the web. In *SIGCOMM*, 2011.

[22] L. M. Smith, L. Zhu, K. Lerman, and Z. Kozareva. The role of social media in the discussion of controversial topics. In *SocialCom*, 2013.

[23] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *ACSAC*, 2010.

[24] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *SocialCom*, 2010.

[25] B. Viswanath, A. Post, K. P. Gummadi, and A. Mislove. An analysis of social network-based sybil defenses. *SIGCOMM Computer Communication Review*, 41(4), 2011.

[26] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao. Sybillimit: A near-optimal social network defense against sybil attacks. In *SP*, 2008.

[27] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. Sybilguard: defending against sybil attacks via social networks. In *SIGCOMM Computer Communication Review*, volume 36, 2006.