2013 International Conference on Social Intelligence and Technology

Mining Social Media: Challenges and Opportunities

Isaac Jones and Huan Liu
Arizona State University
Tempe, AZ
Email: {Isaac.Jones, Huan.Liu}@asu.edu

***Abstract*—The opportunities presented by social networking have led to millions of users flocking to sites like Facebook, Twitter, and Foursquare. Even sites like Amazon have added the ability for users to interact with one another, though it seems tangential to the site's stated purpose. These social networking sites and social networking features generate massive amounts of data that can be used to draw conclusions about social behavior that could previously only be studied using relatively small sample sizes. This unlocks the ability to validate existing social theories, generate new models for how individuals and groups interact, and leverage the power of the crowd, among others.**

I. Introduction

The early 21st century saw an explosion of online social networking activities with the advent of services like MySpace, Facebook, and Twitter. The massive popularity of these services lead to other social networking sites and services that leverage the power of social interaction. Foursquare and Facebook Places, along with some other small services, popularized adding geographic information to social media interactions. Foursquare users check in to locations, post tips for other users, and can earn badges. Foursquare also allows users to become "Mayors" of locations by checking in to that location frequently, adding a competitive element that affects both a user's friends and their non-friends.

Even sites that are not traditionally social networking sites have added user interaction ability in the desire to take advantage of the popularity of social networking. For example, Amazon not only allows users to post reviews, but to evaluate those reviews for their helpfulness. This allows users to interact with each other, and provides Amazon with an automated method for discouraging fake, unhelpful reviews. Amazon also allows users to post their purchases and wish lists to Facebook and Twitter, adding another level of user interaction that bridges platforms.

Just like other sites though, social networking sites also come and go. Nothing demonstrates this better than the migration en masse from MySpace to Facebook as Facebook opened up its availability to more users in the mid-2000s. This move prompted MySpace to launch a re-design of their website in January of this year. Obviously, the recency of the redesign mean that remains to be seen if this re-design is sufficient to attract users back to the site.

Activity on social networks parallels activity in the real world, meaning that a user's behavior online may be a useful indicator of their behavior in the real world, though certainly not every posts all of their activity online. Some users post almost none of their activity online. However, for those that do this means that, with enough data, we can ask many of the same questions and draw many of the same conclusions that social scientists ask and answer, but on a much larger scale. It is important to note that this does not invalidate the work of social scientists, only as a supplement.

Social networks like Facebook and Twitter have an enormous number of users. According to `newsroom.fb.com`, Facebook has over a billion users, more than the population of the entire continent of Europe. If Facebook was its own country, it would be third largest in the world, behind only China and India. Twitter's user count is similar in magnitude. According to analyst group Semiocast[1], the service has over 500 million users. If Twitter was its own country it would also be the third largest. With user numbers in the hundreds of millions, it is inevitable that massive amounts of data will be generated. Twitter reported that 1.38 million posts (called "Tweets") occurred during the three hours of the State of the Union address this year[2]. On Election Day of 2012, users posted more than 31 million tweets about the election alone, reaching a maximum rate of more than 327,000 tweets per minute. The sheer amount of data involved makes thorough analysis of this data impossible using conventional techniques.

To make sense of the massive amount of data generated by users on these services, new techniques are necessary to reduce the massive amount of information generated to a more manageable amount. To complicate the issue, much of social media data is noisy, making the discovery of meaningful signals in that data much more difficult. The techniques used to find information will naturally vary based on the objective of the analysis. We will use some illustrative examples to show some of the research issues tackled by the Data Mining and Machine Learning (DMML) Lab. We will use ongoing projects to demonstrate the methods, challenges, and opportunities involved.

In this work, we focus on the following research issues:

- Information Diffusion: Understanding the patterns that underly memes and virality in social media.
- Privacy and Vulnerability: Understanding how user choices lead to leaking personal information.
- Trust Prediction: Modeling the emerging and evolving way users develop trust relationships online.
- Sentiment Analysis: Automatically extracting the emotional content of social media items.
- User Migration: Understanding when and why users move from one service to another.

[1]Reported by TechCrunch on July 30th, 2012.
[2]According to `blog.twitter.com`

978-0-7695-4998-9/13 $26.00 © 2013 IEEE
DOI 10.1109/SOCIETY.2013.12

- Location-Based Social Networks: Learning about real-world behaviors through social media evidence.
- Tools for Leveraging Social Media: Helping organizations make use of social media data through analysis.

II. Information Diffusion

With the rise of social networking and social media, the average user's ability to consume information from various sources has been tremendously increased. A user does not just have access to the information that he or she can ac quire during his or her free time, but all of the information that all of his or her contacts acquire during their free time. Considering that those people also have access to a similar scope of information, it is easy to see how information can travel incredibly quickly through a social network.

With this massive speed-up in the spread of information, it is easy to see how the concept of *virality* has emerged in the social media landscape. The word "virality" is closely linked to the word "virus", and for good reason. When social media researchers first started exploring virality, the concepts and models that epidemiologists use to model outbreaks of disease were used to model how information spreads in social media. Once it became clear that these models did not accurately predict the spread of information, new models were developed.

Some researchers work on developing global models for predicting the flow of information, and have been successful. In [25], known information propagation pathways are are used to infer the true network structure.

Other work has been done with information dissemination. In particular, Kawk et al. discovered that the height of retweet trees and number of participating users follows a power law [22]. Boyd, Golder, and Lotan studied the factors that might affect how much a tweet is retweeted, and interpreted retweeting as a method of conversation in [6]. It is generally agreed that retweeting and the spread of hashtags indicate information diffusion [8], [24], [43].

These techniques and results mostly focus on the global level of information diffusion. However, as a user of Twitter or Facebook, it is not helpful to know this kind of global information. If a user has information that he or she wishes to spread around a social network it is not useful to know global statistics about retweeting, like the average or maximum length of tweet trees. It is most beneficial for the user to know which of their friends are more likely to pass that information around.

This problem, the problem of identifying *information spreaders*, is one that requires additional attention, leaving many open research questions. The previous prevailing wisdom was that persons of import in social networks, persons whose PageRank [22] scores were high, were also the chief spreaders of information. By analyzing real retweet patterns, a metric was devised for finding the information spreaders in a given user's local network [40]. This metric was not only able to predict future retweeting, but showed that information spreaders are not necessarily important people. In fact, similarity analysis showed that the groups of information spreaders and important people in a given user's network have very little overlap [40]. Even when the top 20 information spreaders are considered among a given user's followers, less than 10% of those users are important people.

In predicting retweets between a given user and his or her friends, it is important to consider what features of a tweet or collection of tweets affect retweeting. Among the features selected, including combining multiple features. it was found that the features that most effectively predict retweeting the best are URLs and Hashtags. In addition, results from the same paper indicate that information spreaders tend not to be frequent retweeters, which could explain why they tend not to be important people in the network.

Unfortunately, not all users of social media are honestly trying to spread information or even show you the latest hilarious YouTube video. Unsurprisingly, social media is increasingly popular as a mechanism for spreading malware and performing cyber-attacks like phishing. Twitter is constantly fighting back against spammers. In April of 2012, Twitter, Inc. filed suit against the makers of tools that enable spammers to target Twitter users[3].

Unfortunately, the administrators of Twitter will never be able to keep up with the onslaught of spam, though they may be able to limit it. In order to protect users from the possible negative impacts of spam on social media, it may be useful to be able to trace the origin of a given piece of information. This problem, called the *information provenance* problem, is especially difficult on social media.

In social media, unlike traditional journalism, sources are rarely cited or acknowledged. There is no requirement that a user on Twitter verifies that the information he or she posts is factually correct before posting it. Twitter, though they take action against spam, does not take action against incorrect information. Because of these things, it would be useful to users if the pathways that information took to reach that user were clear. Knowing the true source of a piece of information allows the user to make decisions about whether or not to trust that information based on the trustability of the actual source, rather than the trustability of the friend that passed on that piece of information.

Tracing the source of a piece of information and/or finding the path it took from a given source to a given destination is the information provenance problem. One possible solution to this problem is, given a graph with a set of known destinations, referred to as *terminals*, algorithmically find the sources, referred to as *root nodes*. Given that this problem is NP-complete, the algorithm proposed makes use of one or both of two hypotheses. These two hypotheses are called the *Degree Propensity* and *Closeness Propensity* hypotheses.

These two hypotheses encode assumptions about how information is likely to travel. The Degree Propensity hypothesis makes the assumption that information is more likely to spread to and from nodes with a high degree. This implies that nodes in the information provenance path have high degrees relative to their neighbors and other potential nodes in the path. The Closeness Propensity hypothesis makes the assumption that sources are close to the terminals. This implies that the path between root nodes and terminals is a short path.

By taking advantage of these hypotheses to create heuristics, the algorithm showed improvement over the algorithms used as the baseline. These reference algorithms were Rumor-centrality [36], Effectors [23], and NetSleuth [34] for a dataset consisting of provenance paths collected from Facebook and provenance paths collected from Twitter.

[3]According to `blog.twitter.com`

III. Privacy and Vulnerability

Tracing the spread of information through social media if useful when a particular piece of information is intentionally seeded on the network and the individual or group of individuals wanted to maximize the spread of information. However, when the information is private or sensitive information that the owner would rather not be spread, the questions shifts from a question of dissemination to one of restriction. In the highly public social media landscape, how does a user keep his or her information from being spread or accessed by others?

This question asks how users maintain their privacy on social networking sites and in social media. The question of privacy on social media has received considerable attention, both in the mainstream media, a cursory news search on Google reveals more than 524 million results for the search term "Facebook privacy"[4], and in the academic community.

By studying privacy settings on Facebook, [14] found that the majority of users keep their privacy settings at the default. Similarly, [26] points out a lack of privacy awareness on social networking sites and in social networking profiles. The researchers found a very large number of profiles where users use a very large vocabulary of terms to describe their passions and interests.

Even without accessing individual personal information, a user's privacy can be at significant risk. Wondracek et al. propose a simple scheme to analyze group membership information that de-anonymizes user information and thus breaches privacy in [42]. Zheleva and Getoor show how adversaries can exploit social network's privacy settings to predict the attributes of users, even those marked as private in [45]. With these great risks to privacy, Krishnamurthy and Wills discussed the problem of personally identifiable information leaking on social media in [19]. They also discussed the how this information can be misused by adversaries.

The risks to privacy and vulnerability to exploitation warrant a response assisting users in protecting their privacy. Fang and LeFevre in [9] focused mostly on changing existing privacy settings to protect information, but this ignores the previously discussed issue of inferring private attributes from public information. Baden et. al. proposed a framework in [3] wherein public/private key pairs dictate availability for private information to different groups of individuals. However, this system is impractical for use in an actual social network, as it substantially increases response times from social networking sites, which may be unacceptable to users.

For this reason, a system was developed using profile and network information to identify the friends in a user's network that expose a given user to a breach of privacy [16]. This operates by categorizing a user's attributes into two groups, *individual attributes* and *community attributes*. These two sets of attributes describe the information that is available about a given user online.

Individual attributes describe information that is unique to that specific user. These attributes include things like gender, birth data, phone number, home address, place of work, etc. These attributes are the attributes that a user is likely most interested in protecting from unauthorized access. It is easy to see how an adversary could cause damage to a user by knowing their home address or phone number.

[4] `www.google.com/search?tbm=nws&q=facebook+privacy`

The risk to a user's privacy through the accessibility of his or her individual attributes is compiled and formulated into an index called the *I-index*, which stands for Individual Index. Since the risk to individual attributes affects only the user whose attributes are possibly being revealed, the I-index essentially measures the risk that a user incurs on his or her own information being accessed by potentially adversarial third parties. The contribution of individual attributes to this I-index is also weighted by the sensitivity of the attribute in question. For example, analysis of a Facebook data set in [16] showed that less than 1% of Facebook users revealed their phone number publicly. This indicates that users who show their phone number are more negligent of privacy settings than those that do not, so a user who reveals their phone number would have a higher I-index than those who do not, or those who reveal less sensitive attributes, like gender.

Community attributes is a group of attributes that describe a given user's friends or friends of friends. Thus, knowing these attributes about an individual reveals, either directly or indirectly, information about the individuals in that user's close network. These attributes include, but are not limited to, a user's friends list, the pictures he or she is tagged in, his or her wall interactions with other users, the groups he or she is a part of, etc. It is less clear how knowing these attributes about a given user can pose a great risk to other users, but it is possible for this information to make inference of other information easier.

The risk to a user's friends from a given user's community attributes is quantified in the *C-index*, which, like the I-index, stands for Community Index. While the I-index is a measure of internal risk, the C-index is a measure of external or projected risk. The Facebook data set analyzed in [16] contained only one community attribute, the friends list. Because of this limited number of attributes, the formulation used for the I-index is not applicable. There must, however, be a way to adjust the index value for each user so that users who make their friends more vulnerable receive higher index scores. This is accomplished by weighting the index score by the number of friends that user is connected with. This weighting means that users who have a lot of friends but still show open their friends to vulnerability receive worse scores than users do the same, but put fewer other users at risk.

These two indices are then combined to create two derived indices, the *P-index* and *V-index*. The P-index is a measure of a user's Publicity. It measures how much information a user has available online and, indirectly, that user's visibility. The V-index is a measure of a user's Vulnerability. It measures how vulnerable a user is to having his or her information leaked online. As expected, this is a function of the user's P-index in combination with that user's friend's P-index, as a user's real vulnerability depends on both the publicity of his information and the publicity of his or her friends' information.

Lastly, the work of [16] presents a methodology for systematically reducing a user's vulnerability. As expected, removing the user's most vulnerable friend decreases the user's vulnerability for every single user in the data set. In addition, removing the two most vulnerable friends also decreases vulnerability for every user in the data set. However, these methods require substantial computation, as the V-index must be recomputed after removing each friend from a users set of friends. Simpler methods, like computing the V-index for all users and then removing the friend with the highest V-index among a users

group of friends results in a decrease in V-index for 95% of users in the Facebook data set [16].

IV. Trust Prediction

As discovered in the work on Privacy and Vulnerability, many users of social media are not concerned with their privacy and the privacy of their friends. For the skeptical user who wishes to keep their information private, this leads to the question of trust in social media. Though Facebook profiles are not analyzed to decide if they are trustworthy or not, researchers do work on trust issues in social media.

With the explosion of social media availability and its increasingly pervasive use in our daily lives, the question of from whom we can accept information and with whom we should share information [13] is increasingly important. In our network, who can we trust to provide reliable information? This is particularly important when applied to e-commerce. Of the hundreds or thousands of reviews posted for a given product, which ones can we trust to show an accurate picture of the product he or she received? Sites like eBay[5] and Epinions[6] that require deep user interaction have trust mechanisms built into the core of their business model.

In the recent past, a lot of research has been done concerning trust online. Recommendation systems have been been developed by [13], [29], and [38] that use trust information as part of the recommendation system. These recommendation systems are called *trust-aware*. Trust has also been factored into systems that search for user-generated content of high quality [17], [28]. Trust relationships have even been factoring into viral marketing applications [35].

However, none of these systems can circumvent the fact that trust information, in the rare cases when it is even available, is very sparse. In addition, the observed number of trust relations follows a power law distribution, with many users showing very few explicit relation and a small minority of user showing a large number. To remedy this issue, the problem of *trust prediction* is proposed.

Trust prediction aims to address the problem of sparseness in trust relationships by inferring trust relationships between users where no explicit relationship exists. Trust prediction has a substantial existing body of work in the active literature. In [5], [15], [27], and [30], other researchers demonstrate systems for trust prediction. However, these methods use an approach that result in highly imbalanced numbers of class labels. This lack of balance makes classification difficult, both in the supervised methods used in [27] and [30] and the unsupervised method used in [5] and [15]. This indicates that additional information is necessary to make an effective prediction system for trust relationships.

The issue of trust does not just exist in the world of social networking. Social scientists have been research trust for many years before the advent of social networking. With this in mind, it makes sense to attempt to utilize existing social science concepts to enhance the ability of trust prediction systems. According to [27], *homophily* is one of the most important social science theories that attempt to explain why individuals decide to trust one another. By factoring the homophily effect into a trust prediction system, it stands to reason that better performance can be achieved.

The homophily effect suggests that users are more likely to trust each other if they are similar. For example, a user who is interested in buying a product on Amazon[7] is more likely to trust a review of that product if they see that the user who posted that review has similar tastes about other items. Exploiting this effect expands the ability to perform trust prediction and enables more research on trust prediction.

However, before the homophily effect can be exploited to enhance trust prediction systems, it must be proven that the homophily effect actually exists and substantially impacts real trust relationships. Though this result may seem intuitive to an informed researcher, it is important to validate every non-trivial assumption made in the process of developing a system or framework. If a non-trivial assumption is not validated, understanding of the fundamental research result is threatened. In the most recent work that deals with trust prediction, [39], this assumption is validated through empirical evidence.

The effect of homophily on trust relationships is quantified as two questions, the first being: Are user with trust relations more similar in terms of their ratings than those without? Here, ratings refers to items rated in data sets collected from product ratings websites Epinions and Ciao[8]. By using cosine similarity between two user's ratings as the measure of similarity of both trust relationships and ratings of products, an objective measurement of the similarity between users with trust relationships and users without trust relationships can be analyzed. Looking at the data, the hypothesis is confirmed with a p-value of $5.12e-18$ and $3.76e-21$ in Epinions and Ciao, respectively [39]. This confirms that users with trust relationships tend to be more similar than those without.

The second question is: Are users with higher similarity more likely to establish trust relations that those with lower similarity? This is similar to the previous question, except it deals primarily with future trust relations, where the first question dealt with existing trust relations. Only the Epinions data set contains time sequence information about trust relationships, so this hypothesis cannot be confirmed in Ciao. The formulation for the verification is similar. We create similarity measures between users and divide into two groups. In this case, the two groups are user with high similarity and low similarity. If the hypothesis is correct, more trust relations should be established with the high similarity group than the low similarity groups. The data confirms this hypothesis with a p-value of $7.59e-59$ [39].

With these two hypotheses confirmed, using the results of the hypotheses to improve a trust prediction system, the original goal of the work, can be performed. By adding a regularization term that exploits homophily to a relatively well-known low-rank matrix factorization model, a system for trust prediction that outperforms baseline methods was created. The exact results can be found in [39]. Outperformed methods include Trust Propagation methods described in [15], Jaccard coefficient based methods, Matrix Factorization methods described in [46], and the low-rank matrix factorization model used as a base for the framework described above.

[5] www.ebay.com
[6] www.epinions.com
[7] www.amazon.com
[8] www.ciao.co.uk

V. Sentiment Analysis

As social networking services grow in popularity, the desire to automatically assess their content for useful information also increases. One of the useful pieces of information that can be extracted from social media is sentiment. Merriam-Webster[9] defines sentiment as "an attitude, thought, or judgement propted by feeling." In social media mining, sentiment analysis is then the automated extraction of emotional content from social media data.

There are many reasons why an individual or organization may want to do this. For example, a company may want to assess the public's feelings toward their products by inspecting social media data. Relief and recovery organizations may wish to monitor the sentiment of a population before, during, and immediately after crisis and recovery operations to ensure that the recovery operations were successful and aid recovery efforts if there are still areas that require resources.

However, the data available on social media is very noisy and and generally short-form, which presents substantial challenges for sentiment analysis. Movie reviews and product reviews have been extensively studied in the field of sentiment analysis [31]. However, social media differs substantially from media like movie and product reviews. Firstly, reviews tend to be longer that social media. On Twitter, posts are limited to 140 characters, which limits posts to one or two sentences at maximum. According to Twitter employee Isaac Hepworth[10], the average length of a Tweet is approximately 30 characters long. This extremely short length obvious makes sentiment analysis much more difficult, especially when compared to relatively long texts of a movie or product review.

Secondly, users of social media often improvise words or use phrases to mean things that were not originally intended. It is very rare to see improvised words in formal reviews, but using improvised words like "OMG" as an exclamation or using abbreviated words like "till" to mean "until." Existing systems often rely on pre-defined vocabularies, which fail to capture these improvised words [41]. In addition, high-level linguistic concepts like sarcasm are usually omitted from reviews, but since social media is a less formal a more conversation medium, it is relatively common to a construction that uses sarcasm or irony humorously. However, automated systems have a difficult time processing these linguistic peculiarities, so these may be missed by these systems.

Lastly, not all social media posts have sentiment attached to them. A review for a movie or product is by its nature designed to be either supportive or unfavorable toward the subject of review. Social media posts do not necessarily have any sentiment at all attached to them. For example, the tweet "Dinner at my house tonight at 6:00pm." has no sentiment attached to it. It is simply a user telling his friends what time dinner was served or will be served. This adds a third class to the sentiment analysis task, the neutral class.

One advantage that social media has that traditional media does not is that there are links between the users that post social media items. Relationships between users may provide hints as to the message's semantic content. Thus, social tie information and the social media information can be combined to predict sentiment more accurately than with the media information alone. The idea of combining social ties and individual post information is not new, Tan et. al. did this in [37]. However, the work of Tan et. al. operated at a user-level, not the message-level that is more meaningful for social media. An overview of the work in [18] is presented here, but full details can be found in the original work.

In order to analyze social media on a more granular level, each piece of social media is thought of as a message, and represented as a term-frequency vector. This allows for a concise matrix representation of the entire data set, which is in this case comprised of Tweets collected during important events. This matrix is annotated with an additional matrix that represents the sentiment content of each term. Techniques that do not take social ties into account when determining the sentiment of a message do not need any more representation. However, the work done does take social ties into account in an attempt to mitigate some of the challenges posed by the nature of social media messages as described above. To add social tie information into the representation, it is necessary to create a third and fourth matrix linking each message to a user and linking users to other users, respectively. With these additional matrices, additional information can be utilized in the problem formulation that allows for the incorporation of social tie information.

Before checking if adding social ties information adds real information to the problem of sentiment analysis, it is important to first validate the assumptions that underlie the intuition. Two such assumptions were validated, the first being the *sentiment consistency* theory. Sentiment consistency states that the sentiments of two posts from a given user are more likely to be the same than the sentiments of two random posts. In [1], Abelson discusses consistency in sentiment, but this theory was not previously validated on social media data. This theory was checked against two datasets, the first being the Stanford Twitter Sentiment (STS) dataset[11] and the second being the Obama-McCain Debate dataset[12]. [18] finds that there is sufficient evidence to confirm this theory in both data sets with $\alpha = 0.01$.

The second assumption that this model makes is the *emotional contagion* theory. This theory states that the sentiments of two messages posted by friends is more likely to be the same than two random messages. Again, analysis on both of the data sets in listed above performed in [18] shows that sufficient evidence exists to confirm the theory with $\alpha = 0.01$.

Knowing that these two assumptions hold for sentiment in social media messages, a model can be created that takes this information into account. This model, called *SANT* and explained in detail in [18], incorporates social tie information, the sentiment consistency effect, and the emotional contagion effect. However, due to the representation of social media messages as a matrix, *SANT* requires an additional performance enhancement to reach peak performance. Because Twitter data has so few words per individual message and so many message, the vocabulary involved in Twitter data is very large. This mean that the matrix representing messages as term-frequency vectors is an extremely sparse matrix. To combat this effect, a sparsity regularization parameter is incorporated into the model. This costs the model some amount of computation time, but increases the accuracy of the model.

[9]From Merriam-Webster Online at `www.merriam-webster.com`

[10]`twitter.com/isaach`

[11]Available at `www.standford.edu/ alecmgo/cs224n/`

[12]Available at `bitbucket.org/speriosu/updown/src/5de483437466/data/`

The resultant model outperforms the baseline models, which in this case are least square using only sentiment relation information; least squares using both tweet content and sentiment relation information; Lasso, a sparse formulation of least squares, with only sentiment relation information; and Lasso using both sentiment relation information and tweet content. The *SANT* model reaches more than 75% on polar sentiment classification (only positive or negative) and more than 55% accuracy on three-class sentiment classification (positive, negative, and neutral).

VI. User Migration

The incredible growth in the usage of social media in the past decade has been accompanied by a growth in the number of operating social media sites. Though users may want to get the most fulfilling experience out of every site, they are constrained by limited time and attention. Since users cannot stay engaged in every social media site, their attention must wander from site to site. This dynamism is encapsulated by the user migration problem. This problem seeks to understand how users select which social media site on which to spend their limited attention resources.

Understand this can help the owners of social media sites curate their sites in such a way to retain the users already present on the site as well as attract new users. The work in [20] demonstrated not only that user migration is a valuable problem to study, but that it can be studied in a meaningful way, user migration has identifiable patterns, and that it is possible to influence those patterns.

In order to study user migration patterns, the types of migration patterns that are likely to exist among social network populations must be defined. These two types of migration are defined as *Site Migration* and *Attention Migration* by [20]. Site Migration describes the type of migration where users of sites are mutually exclusive. This means that a user of site one is not a user of site two. This can happen when a user creates an account on one site after deleting or deactivating their account on another site. For example, a user trying to promote his or her music may have deleted their MySpace account when they created a Facebook account.

The other type of migration, Attention Migration, is measured by a user's activity on two sites. In this type of migration, the accounts are kept active on both sites, but activity decreases, possibly sharply, on one site while it increases on the other. In this type of migration, the user in the example above did not go so far as to delete their MySpace account but instead stopped logging in to MySpace and stopped updating information in their profile, preferring to perform these actions on their Facebook account instead.

Attention migration requires the definition of another measure for determining if and when a migration happened. This measure is *User Activity*. This is a binary measure, indicating that the user is either active or inactive, but not somewhere in between. A user is considered active if he or she has performed at least one action on the site in the last time interval δ. Conversely, a user in considered to be inactive if he or she has not performed an activity in that time interval.

Previous work described herein has primarily used the data set of only one social network in data analysis. Obvious, this is not sufficient for considering user migration patterns. For the work of [20], seven different social networking sites were considered, Delicious[13], Digg[14], Flickr[15], Reddit[16], StumbleUpon[17], Twitter, and YouTube[18]. However, naive collection of these data sets does not address the problem of resolving user identities across multiple sites. This problem is address by Zafarani and Liu in [44]. This problem was avoided in the collection of the data set for [20] by taking advantage of BlogCatalog[19], which allowed for the collecting of user profiles on all seven sites knowing that the profiles represented the same user across all seven sites.

Since user migration is necessarily time-dependent, it is not sufficient to just collect a set of user profiles from each website once. Data must be collected multiple times over a period of time to accurately capture migration patterns. For this reason, user profiles on BlogCatalog were collected three times, with each collection one month apart. This sets the δ for determining user activity at one month, as that is the most granular possible resolution on this data set. This divides the activity into two time periods, called Phase 1 and Phase 2. Since the assumption cannot be made that users were active before data collection began, the only measure of activity is between two data collection times. Since there are three data collections times, there are two periods of measurable activity.

Initial analysis of the data substantiates the claim that attention migration exists between social networking sites. One of the most significant migration patterns noted in [20] indicates that a substantial number of users (16%) migrated from Reddit to StumbleUpon and Digg. In addition, it is noted that there is a significant quantity of mutual migration, that is migration in both directions, between StumbleUpon and Delicious. Twitter and StumbleUpon also have a substantial quantity of attention migration to the sites from all locations.

This observed effects could be caused by many things. To ensure that this is caused by a time-dependent process like attention migration, a statistical test must be used that demonstrates that effect of the sequence of events in time significantly affects the process. One of the most commonly used tests for this is the *shuffle test* described in [2]. The objective of performing the shuffle test in this case is to prove that user activity on a site predicts user migration.

Since activity is determined on a site-by-site basis, the shuffle test also operates on a site-by-site basis. Thus, this migration experiment is actually divided into seven different migration experiments, one for each of Delicious, Digg, Flickr, Reddit, StumbleUpon, Twitter, and YouTube. The results of the shuffle test performed in [20] show that migration is actually only statistically significant (p-value ≤ 0.05) for StumbleUpon, Twitter, and YouTube. Though not reaching the level of significance, the data for Flickr was substantially different before and after the shuffle test was performed, which prompted a further investigation into the results from the data set. It was found that the Flickr data set was quite small, which prevented a statistically-oriented test from reaching conclusions. Further investigation into user migration specifically on Flickr may be warranted to determine if user migration is

[13] www.delicious.com
[14] www.digg.com
[15] www.flickr.com
[16] www.reddit.com
[17] www.stumbleupon.com
[18] www.youtube.com
[19] www.blogcatalog.com

significant for that site.

Though this study showed that user migration is significant for some social networking sites, it left some questions open for future work. In particular, no effort was made to study user migration as a result of site changes. With the redesign of MySpace recently published as mentioned in Section I, there is a great amount of potential research to be done on how changes to sites affect user migration and user attention patterns.

VII. Location-Based Social Networks

With the increasing popularity of networks like Foursquare that integrate geographic data with social networking abilities, it stands to reason that these networks would be excellent sources of data to analyze. Recent surveys have shown that approximately 4% of people living in the United States use some kind of location-based social networking service, be it Foursquare[20]. Gowalla[21], or Facebook Places[22]. Zickuhr and Smith discovered that approximately 1% of Internet users take advantage of these services daily [47].

In this new environment, people share their activities in new ways. The idea of "checking in" does not exist in other social networks. In Location-Based Social Networks (LBSNs), checking in with another person from a user's friend group is an action that can be verified by inspecting the friend's check-in history, unlike a Facebook status update, where users can be arbitrarily tagged. By exploiting these new types of activities, more refined systems can be developed that benefit the users of these networks, like location recommendation [4]. By utilizing this new kind of data, there are even opportunities in areas like disaster relief [10].

The obvious benefit of having a user's check-in history is that a researcher has a listing of all of the places a given user has visited. However, there is benefit in considering both the location the user checked in at and the time at which the user performed this action. There may be a pattern is the user's data that access to the timestamps allows visibility into that would otherwise be lost. For example, a user may go to a coffee shop every week day at approximately 2:30 pm. When attempting to perform location prediction, knowing that pattern is very important.

However, this is not the only factor. Social correlation theory suggests that it may be valuable to consider a user's social ties, as a user's behavior may be strongly informed by his or her friend's behaviors [2]. For example, two coworkers may arrive at their place of work at the same time, go to lunch together, and leave at about the same time. Two friends who live in disparate parts of the country will probably go to a lot of the same places if one ever visits the other.

By combining these two factors, social ties and historical analysis, a system can be formulated that outperforms systems using these two pieces of information separately. Before considering the methodology for combining the two factors, it is important to first important to consider what unique features of each set of data contribute information to the problem of predicting a user's next location.

The historical information present in a user's check-in history has two interesting and important properties. The first of these properties is that a user's check in history follows a power law distribution. When inspecting aggregate number of check ins, users have a tendency to check into a few places many times and many places only a few times. This is intuitively supported by considering the example of an employee who checks in at work every week day, but goes to a different restaurant for lunch each day. In an average weekday, the office or place of employment accrues five check ins, where each restaurant accrues only one. The second property is time-dependance. Many check ins occur at consistent times of the day or consistent cycles in a time period. In addition, many check ins occur in a predictable sequence. Considering the example of the employee above, a check in at lunch always occurs after a check in at the place of employment and re-checking in at the place of employment follows the lunch check-in.

The social information present in a user's network information also has important properties. Using a data set collected from Foursquare, it was found that pairs of users who are friends have three times more shared check ins that pairs of users who are not friends [11]. This indicates that the friendship relationship between users strongly influence a given user's future check in locations. To confirm this intuition, [11] performed a verification test, and found that the check in similarity between friends is greater than the check in similarity between random users with p-value $2.6e-6$.

Knowing these properties about check in histories, it is necessary to pick a model that can best take advantage of the indicated patterns. In [11], researchers selected a Pitman-Yor model [32], [33]. The Pitman-Yor model was originally designed to analyze text documents, however the historical properties of a user's check ins show strong similarities to a document model. In the basic Pitman-Yor model, there is an assumption that there is a corpus of documents to be analyzed. In LBSN modeling, this corresponds to a collection of check ins from multiple users. Individual documents in the collection correspond to an individual user's check in history. Within the document level, the paragraph structure of a document can be viewed as a user's check in history on a month-by-month basis. Sentences can be viewed as weekly check in histories. Phrases can be viewed as daily check in histories. Finally, individual check ins correspond to individual words. In [48], Zipf finds that word frequency in document collections follows a power law, which matches the finding in [11] that check in histories also follow a power law. These correspondences and the matching of the power law distributions indicate that the Pitman-Yor model is a good model for predicting check ins.

Next, we must integrate the social tie information into the model. This is done by adding a weighted regularizer to the predicitons generated by the Pitman-Yor model. This allows researchers working with the model to adjust the effects of the social ties and historical information. The social tie regularizer consists of the predicted probability of the two users checking in together multiplied by the similarity of the two users, assuming that the two users are friends.

By using this model, called the Social-Historical Model, the results outperformed the baseline models for predicting user locations. The baseline models used in [11] were the Most Frequent Check In Model [7]; Most Frequent Time Model, a model that predicts based on the place most frequently checked in at the current time; and the Order-k

[20] www.foursquare.com

[21] gowalla.com

[22] www.facebook.com/facebookplaces

Markov Model, which considers the context of the latest k check ins to predict the next check in. It is worth noting that the first model, Most Frequent Check In, is simply an Order-0 Markov Model. Though the Social-Historical Model outperforms all the baselines, valuable information can still be gained by varying the weighting parameter.

To compare the impacts of weighting the two contributing factors, the weights were varied from 0 (only social information) to 1 (only historical information). At one extreme, 0, the performance of the model is always the worst. This indicates that social information is not enough to obtain good location prediction. At the other extreme, 1, the performance of the model is not the worst, but can be improved upon. This suggests that social tie information is important, but not overwhelmingly so. The optimal performance is achieved when the weighting factor is approximately 0.7, suggesting that historical ties are significantly, but not overwhelmingly, more important than social ties.

VIII. Tools for Leveraging Social Media

The researchers at the DMML Lab are committed to developing tools that assist outside organizations in using social media to further their own goals in addition to doing meaningful research. Even the activities with research results as the primary objective are strongly motivated by problems that users and organizations using social media face when trying to make sense of the massive amount of data available. To this end, a variety of tools have been created, two of which will be discussed here, that expose the results of research to organizations, primarily those in the Humanitarian Aid and Disaster Relief (HADR) area.

A. TweetTracker

Chief among the tools, and the basis for the other tool discussed here, is TweetTracker[23]. This system allows organizations to monitor Twitter data in real-time using simple and intuitive interfaces. TweetTracker obtains Twitter data by monitoring the Twitter Streaming API. The streaming API provides approximately 10% of the Tweets published to Twitter in a easily-accessible format. Twitter states that the streaming API "Returns a small random sample of all public statuses."[24]. This means that the most current conversations and trending topics are likely represented in the random sample, due to the sheer volume of tweets published on those topics.

In using TweetTracker, one of the first steps a user takes is to define an event. This event is a collection of keywords, geographic bounding boxes, and user accounts. TweetTracker stores this information and applies this as a filter to all of the incoming data from the streaming API. For each Tweet the system receives through the API, the contents of the Tweet are compared against the keywords, the geographic origin is compared against the bounding boxes, and the originating user is compared against the user accounts. If the tweet is found to contain any of the listed keywords, originate within any bounding box, or originate from any of the specified users it is stored in a database for future analysis.

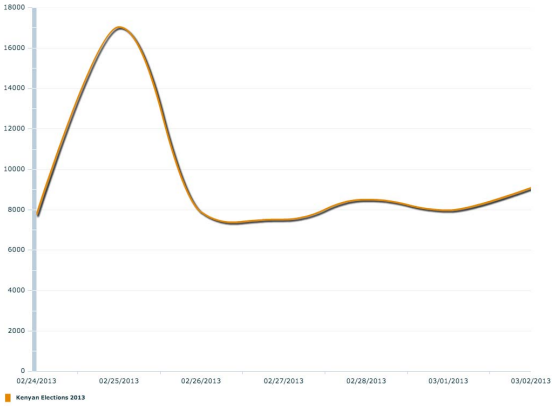

Fig. 1. An example of Keyword Frequency trending on 2013 Kenyan Elections data

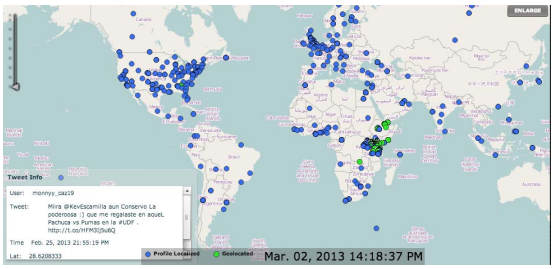

Fig. 2. An example of Twitter data plotted on a world map

After defining an event and allowing the system to collect data, TweetTracker's analytics capability can be leveraged to make sense of the data. The first step is to determine which terms in the data set have the most interesting behavior in the data set collected. This can be accomplished by using the TweetTrends section of TweetTracker, which shows users keyword frequency trends over time. After selecting an event, setting a resolution, and defining a time window of data to look at, the user is presented a chart such as the one shown in Figure 1. This allows the user to look for periods of time that warrant further inspection by virtue of their term frequency behavior. In the example shown, the February 25th date may warrant further inspection because of the substantial spike of Tweet volume observed on that day.

After identifying an interesting region of time, a user of TweetTracker proceeds to the Tweetalyzer section of TweetTracker, which allows users to inspect Tweets in more detail. First, all of the Tweets in the selected time region are plotted on a map, according to their geographic location. An example of this map for the February 25th date is shown in Figure 2. Each dot on the map represents a tweet, which allows users to see where each tweet is coming from. This is useful for disaster relief efforts, as precise geolocations, represented by green dots, give an exact picture of where users Tweeting in the hopes for relief aid are located. Blue dots use the user's profile to determine the location of the tweet, so this metric is less accurate as it relies on inferred data.

In addition to the geographic map, TweetTracker generates and presents a Tag Cloud, an example of which can be seen in Figure 3. This tag cloud demonstrates what Twitter users

[23] An overview of TweetTracker can also be found at `tweettracker.fulton.asu.edu`.

[24] According to `dev.twitter.com/docs/api/1.1/get/statuses/sample`

Fig. 3. A Tag Cloud generated from the tweets plotted on the map in Figure 2

are talking about that may be related to the original topic of inquiry. These could be subtopics, in the Figure 3 the name of the Kenyan politician Uhuru Kenyatta appears, or related topics; in the same figure "debate" appears, possibly referring to the debate between Presidential election candidates that occurred in Kenya on the 25th of February.

After analysis is done in TweetTracker, users can use the Search/Export feature to obtain data from TweetTracker for further analysis. This section of TweetTracker allows users to export Tweets in a tab-separated value format, XML format, or a format suitable for analysis in other tools. One such tool, developed internally, is TweetExplorer[25], which allows for further analysis and is focused on network information. TweetTracker is currently in use by humanitarian organizations like Humanity Road, who uses the tool to gain first-hand knowledge and maintain situational awareness during times of crisis. More information about TweetTracker can be found in [21]. It is important to note that TweetTracker can be used by any agency to study whatever is needed on Twitter. TweetTracker does not impose any assumptions on what a user might want to find from Twitter data. This gives TweetTracker great flexibility in the analysis tasks it can take on.

B. ASU Coordination Tracker

The second tool under active development is the ASU Coordination Tracker (ACT). This tool aggregates social media data as well as user-submitted data to assist humanitarian agencies coordinate disaster relief efforts across multiple agencies. Noting that an effective tool did not previously exist to coordinate disaster relief efforts across multiple agencies, researchers created a system that could fill this void. This system is comprised of three modules; a crowdsourcing module, a small group module, and an analytics module. The bulk of the work is performed in the small group module, but all three modules perform important functions.

[25] tweetracker.fulton.asu.edu/tweetexplorer

The first two modules, the crowdsourcing and small group modules, are data input modules. They accumulate resource requests from individuals in the crisis situations that require attention. The small group module is fairly simple. It allows groups of responders or private citizens to submit requests for individual aid. This aid can take any form, the request description is free-form. The crowdsourcing module uses TweetTracker to automatically collect resource requests from publicly available social media data. Obviously, due to the high amount of noise in social media data, requests generated from this module are subject to a higher level of scrutiny that requests originating from the small group module.

After submission, all requests for resources go into a queue that allows responding organizations to assess the magnitude of the request, the urgency of the request, and the availability of resources to satisfy that request. After an agency has selected a resource request to respond to, that request is removed from the active request queue. By allowing agencies to select which requests they respond to and then preventing other agencies from responding to that same request, the ASU Coordination Tracker reduces the redundant responses and waste of resources that often result when more than one agency responds to the same request.

Once a request is taken on by an agency, the request is marked as in progress, as distinct from satisfied or completed. If the responding agency does not, for some reason, actually satisfy the request, this allow the request to be automatically moved back into the active queue without needing the requester to publish an additional request to the system. Again, this prevents a waste of relief resources by reducing the likelihood that a requester will submit a duplicate request and thus waste otherwise useful resources. In addition, this adds accountability to an agency, as an agency which fails to respond to many of the requests they claim to satisfy will not be trusted to satisfy their requests in the future.

The last step and the last module is the analytics module. Response coordinators may desire to understand the progress of the relief efforts in statistical terms, as that allows coordinators to know if they need to request more resources from an agency's home offices, for example. ACT provides these statistics and other analytics in the analytics module, which provides statistics about current request fulfillment status, spatial and temporal distribution of requests, and distribution and contribution of each responding organization. For example, response coordinators may decide to request additional response resources if the temporal distribution of requests indicates that requests for resources are not slowing down.

ACT's ability to coordinate disaster relief was tested at the ASU Crisis Response Game, a simulated disaster that tested the ability of social media to provide valuable information in times of crisis. More about ACT can be found in [12].

IX. Conclusion

In this paper, we have presented a number of areas of active research in social media mining. This work covers may areas, including, Information Diffusion, Privacy and Vulnerability, Trust Prediction, Sentiment Analysis, User Migration, and Location-Based Social Networks. In addition, a brief discussion of some of the tools available to outside organizations interested in learning from social media and using social media for disaster relief and coordination has been presented. This is

not an exhaustive list of topics of interest in social media, and future work on all of these topics and more is in progress.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions from the members of the DMML Lab, particularly those who contributed to the work discussed here. The work described herein is, in part, supported by ARO (#025071), ONR (N00014-11-1-0527 and N00014-10-1-0091), and NSF (#IIS-1217466).

REFERENCES

[1] R. Abelson. Whatever Became of Consistency Theory? In *Personality and Social Psychology Bulletin*, 1983.

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and Correlation in Social Networks. In *KDD*, 2008.

[3] R. Baden, A. Bender, N. Spring, B. Bhattacharjee, and D. Starin. Persona: An online social network with user-defined privacy. *ACM SIGCOMM Computer Communication Review*, 39(4):135-146, 2009

[4] P. Barwise and C. Strong. Permission-Based Mobile Advertising. *Journal of Interactive MArketing*, 2002

[5] P. Borzymek, M. Sydow, and A. Wierzbicki. Enriching trust prediction model in social network with user rating similarity. In *International Conference on Computational Aspects of Social Networks*, 2009.

[6] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 43rd Hawaii International Conference on Social Systems*, 2010.

[7] J. Chang and E. Sun. Location 3: How Users Share and Respond to Location-Based Data on Social Networking Sites. In *ICWSM*, 2011.

[8] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. A. Goncalves, and F. Benevenuto. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *Proceedings of the Workshop on Languages in Social Media*, 2011.

[9] L. Fang and K. LeFevre. Privacy wizards for social networking sites. In *The 19th International World Wide Web Conference (WWW)*, 2010.

[10] H. Gao, G. Barbier, and R. Goolsby. Harnessing the Crowdsourcing Power of Social Media for Disaster Relief. *IEEE Intelligent Systems*,2011

[11] H. Gao, J. Tang, and H. Liu. Exploring Social-Historical Ties on Location-Based Social Networks. In *The Sixth International AAAI Conference on Weblogs and Social Media (ICWSM2012)*, 2012.

[12] H. Gao, X. Wang, G. Barbier, and H. Liu. Promoting Coordination for Disaster Relief - From Crowdsourcing to Coordination. In *SBP*, 2011.

[13] J. Golbeck. Generating predictive movie recommendations from trust in social networks. In *Trust Management*, 2006

[14] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *The ACM Workshop on Privacy in the Electronic Society*, pages 71-80. ACM, 2005.

[15] R. Guha, R. Kumar, P. Ragavan, and A. Tomkins. Propagation of trust and distrust. In *WWW*, 2004.

[16] P. Gundecha, G. Barbier, and H. Liu. Exploiting Vulnerability to Secure User Privacy on a Social Networking Site, In *The 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2011.

[17] Y. Hu, A. John, F. Wang,and S. Kambhampati. ET-LDA: Joint Topic Modeling for Aligning Events and their Twitter Feedback. In *AAAI*, 2012

[18] X. Hu, L. Tang, J. Tang, and H. Liu. Exploiting Social Relations for Sentiment Analysis in Microblogging. In *Proceedings of the 6th International Conference on Web Search and Data Mining*, 2013.

[19] B. Krishnamurthy and C. Wills. On the leakage of personally identifiable information via online social networks. *ACM SIGCOMM Computer Communication Review*, 40(1):112-117, 2010.

[20] S. Kumar, R. Zafarani, and H. Liu. Understanding User Migration Patterns in Social Media. In *The Special Track on AI and the Web at the Twenty-Fifth AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011

[21] S. Kumar, G. Barbier, M. A. Abbasi, and H. Liu. TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief. Demo in *The 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011

[22] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International World Wide Web Conference*, 2010.

[23] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila. Finding Effectors in Social Networks. In *Proceedings of the 16th ACM SIGKDD*, 2010.

[24] K. Lerman and R. Ghosh. Information contaigon: an empirical study of the spread of news on digg and twitter social networks. In *Proceedings of Fourth International Conference on Weblogs and Social Media*, 2010.

[25] J. Leskovec, M. Gomez-Rodriguez, and B. Schoelkopf. Structure and Dynamics of Information Pathways in Online Media. In *WSDM*, 2013.

[26] H. Liu and P. Maes. Interestmap: Harvesting social network profiles for recommendations. *Beyond Personalization*, 2005.

[27] H. Liu, E. Lim, H. Lauw, M. Le, A. Sun, J. Srivastava, and Y. Kim. Predicting trusts among users of online communities: an epinions case study. In *EC*, 2008.

[28] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *WWW*, 2010.

[29] H. Ma, D. Zhou, C. Liu, M. Lyu, and I. King. Recommender systems with social regularization. In *WSDM*, 2011.

[30] V. Nguyen, E. Lim, J. Jiang, and A. Sun. To trust or not to trust? predicting online trusts using trust antecedent framework. In *ICDM*, 2009.

[31] B. Pang and L. Lee. Opinion Mining and Sentiment Analysis. In *Foundations and Trends in Information Retrieval*, 2008

[32] J. Pitman and M. Yor. The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator. In *The Annals of Probability*, 1997.

[33] J. Pitman. *Combinatorial Stochastic Processes*, volume 1875, 2006.

[34] B. Prakash, J. Vrekeen, and C. Faloutsos. Spotting Culprits in Epidemics: How many and Which ones? In *Proceedings of the 12th IEEE ICDM*, 2012.

[35] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD*, 2002

[36] D. Shah and T. Zaman. Rumors in a Network: Who's the Culprit? In *IEEE Transactions on Information Theory*, 2011.

[37] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-Level Sentiment Analysis Incorporating Social Networks. In *SIGKDD*, 2011.

[38] J. Tang, H. Gao, and H. Liu. Discerning multi-faceted trust in a connected world. In *WSDM*, 2012.

[39] J. Tang, H. Gao, X. Hu, and H. Liu. Exploiting Homophily Effect for Trust Prediction. In *The Sixth ACM International Conference on Web Search and Data Mining*. February 4-8, 2013. Rome, Italy.

[40] X. Wang, H. Liu, P. Zhang, and B. Li. Identifying Informaiton Spreaders in Twitter Follower Networks. Technical Report.

[41] J. Wiebe, T. Wilson, and C. Cardie. Annotating Expressions of Opinions and Emotions in Language. In *Language Resources and Evaluation*, 2005.

[42] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel. A practical attack to de-anonymize social network users. In *The 31st IEEE Symposium on Security and Privacy*, 2010.

[43] J. Yang and S. Counts. Predicting the Speed, Scale, and Range of Informaiton Diffusion in Twitter. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010

[44] R. Zafarani and H. Liu. Connecting Corresponding Identities Across Communities. In *ICWSM*, 2009.

[45] E. Zheleva and L. Getoor. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *The 18th International World Wide Web Conference (WWW)*, 2009.

[46] S. Zhu, K. Yu, Y. Chi, and Y. Gong. Combining content and link for classification using matrix factorization. In *SIGIR*, 2007.

[47] K. Zickhur and A. Smith. 4% of Online Americans use Location-Based Services. *Pew Internet & American Life Project*, 2010.

[48] G. Zipf. Selective Studies and the Principle of Relative Frequency in Language. In *Human Behavior and the Principle of Least-Effort*, 1949.