Why do People Retweet? Anti-Homophily Wins the Day!

Sofus A. Macskassy and Matthew Michelson*
Fetch Technologies, 841 Apollo Street, El Segundo, CA 90245
sofmac@fetch.com, matt.michelson@gmail.com

Abstract

Twitter and other microblogs have rapidly become a significant means by which people communicate with the world and each other in near realtime. There has been a large number of studies surrounding these social media, focusing on areas such as information spread, various centrality measures, topic detection and more. However, one area which has not received much attention is trying to better understand what information is being spread and why it is being spread. This work looks to get a better understanding of what makes people spread information in tweets or microblogs through the use of retweeting. Several retweet behavior models are presented and evaluated on a Twitter data set consisting of over 768,000 tweets gathered from monitoring over 30,000 users for a period of one month. We evaluate the proposed models against each user and show how people use different retweet behavior models. For example, we find that although users in the majority of cases do not retweet information on topics that they themselves Tweet about as or from people who are "like them" (hence anti-homophily), we do find that models which do take homophily, or similarity, into account fits the observed retweet behaviors much better than other more general models which do not take this into account. We further find that, not surprisingly, people's retweeting behavior is better explained through multiple different models rather than one model.

1 Motivation

The use of "micro-blogging" services, such as Twitter, has exploded exponentially in recent years. For example, currently, millions of Twitter users post millions of 140-character messages, called "Tweets," about topics ranging from daily activities, to opinions, to links to funny pictures. Beyond the large collection of user generated text, Twitter also has a social network aspect, allowing users to publicly message one another directly, and set up a social network of people who follow one another's Tweets. This rich relational and textual setting has spurred research in a number of areas (beyond traditional network analysis (e.g., (Kwak et al. 2010; Krishnamurthy, Gill, and Arlitt 2008)). For instance, Twitter has been analyzed to discover breaking news (Sankaranarayanan et al. 2009), as a forum for analyzing media events (Shamma, Kennedy, and Churchill 2009), as a vehicle for information diffusion (Leskovec et al. 2007; Lerman and Ghosh 2010; Lerman and Hogg 2010), as a mechanism for language learning (Borau et al. 2009), and even for detecting natural disasters in real-time (Sakaki, Okazaki, and Matsuo 2010).

Much recent work in microblogs as described above tend to treat the social media streams and underlying social networks as large global phenomena where global processes, metrics and statistics rule the day. In other words, the streams, people and links in these social media are all treated as a large homogeneous mass. While such a high-level view of the world is of tremendous use in order to understand large global behaviors, it unfortunately is not appropriate for fine-grained analysis of local behaviors. For example, community detection fails to find meaningful clusters on these large networks (see, e.g., (Leskovec et al. 2008)), information diffusion and other metrics match on macro-level but fails to fit observed data at the micro-level.

We argue in this paper that context is critically important when one wants to delve into the details. Not all links area created equal, not all people are the same, and not all pieces of content are interesting. If one can tag people, links and content with semantically meaningful categories, then one ought to be able to generate much finer-grained behavioral and predictive models to understand the dynamics of these social media networks. In particular we here try to understand the contextual factors which makes a person retweet a particular piece of information.

We here take a first step towards being able to characterize context, and identifying processes which lead to people diffusing information to their local network. While there are many types of context one can use, we here focus on generating a profile of "topics of interest" for a user based on past content posted, and then use this profile to gain insight into what makes people propagate information through different behavior models. We show that our profile-based models have a better fit to observed information propagation than a more general behavior model.

The key to our contribution lies in building and using these user profiles. This is done through automatic tagging of people and content into semantically meaningful categories and then using these categories to develop context-specific behavioral models for information propagation. Our approach further relies on being able to match and disambiguate *entities* mentioned in content so that we can track what a person writes about over time. For example, rather

*Work performed while at Fetch Technologies.

than track that a person writes about "Obama" and "Bush" and "Clinton", we would like to learn that repeated instances about "Bush" is likely the president of the United States and that the topic really is Presidents and politics rather than these keywords. We do this by mapping found entities into an ontology, as we describe below, and then keeping track of which ontological concepts show up repeatedly in a user's content. These repeated concepts can then be used as that person's "topic of interest profile", which we can use to map against other content, specifically with respect to what that person decides to propagate.

Once we have these richer context-specific tags, we explore different information propagation behavior-models to get a better idea about what information users tend to "pass on" or propagate. We explore four particular information diffusion models at the individual level in the domain of Twitter, where we have been monitoring 30,000 Twitterers for over a month and have collected over 768,000 tweets from these Twitterers. These Twitterers were identified through a focused sampling to ensure we got a sample set of Twitterers who were aware of each other. We show that our two context-specific retweet behavior-models are better at explaining the observed retweet behavior than a generic or network-based model.

The rest of the paper is outlined as follows: in Section 2 we discuss related work. We then, in Section 3 describe our approach for tagging content and building user profiles, followed in Section 4 by a description of our four behavior-based information-propagation models. Section 5 describes our case study on Twitter data, where we show that our context-models are better than general models. We finish in Section 6 with a discussion of our findings.

2 Related Work

Information diffusion is a topic which is receiving an increasing amount of attention in many areas, social media as well (see, e.g., (Leskovec et al. 2007; Gomez-Rodriguez, Leskovec, and Krause 2010; Sadikov et al. 2011; Lerman and Ghosh 2010; Suh et al. 2010)). Most of these endeavors, however, are focused on developing global statistics and metrics, such as understanding information cascades or learning pathways that information propagated. None of the methods are seriously considering treating individuals differently or relations differently beyond was is captured by the high-level statistics. This short-coming is exactly what we are trying to address in this paper.

Community detection algorithms have received significant attention in recent years (see, e.g., (Clauset, Newman, and Moore 2004; Muff, Rao, and Caflisch 2005; Newman 2005; White and Smyth 2005; Leskovec et al. 2008; Porter, Onnela, and Mucha 2009)). The most common approaches take a graph (such as a social network) and split it into k disjoint clusters, where each cluster supposedly represents a "community" in that graph. This type of approach is appropriate when one can reasonably expect that there is a clear enough signal in the graph, such that the found communities are likely to represent real sub-communities. This is often the case in relatively well-defined, generally small networks, with a single well-defined and appropriate semantic interpretation to the edges.

Being able to tag content and profile users is an area which has begun to receive attention from researchers in social media analytics. Of interest to this paper are two approaches which address different problems using Twitter data.

Chen, et al., 2010 explore the problem of recommending content (Tweets). They build a number of recommender approaches, one of which is "topic" based. They model the topics of a user as a bag-of-words generated from the user's Tweets (with TF/IDF weights). They then compare this feature vector modeling of the topics to a similar feature vector of an incoming Tweet to determine if it should be recommended to the user.

Another approach to analyzing Twitter that uses topics is TwitterRank, which aims to identify influential micro-bloggers (Weng et al. 2010). This approach leverages LDA by creating a single document from all of a user's Tweets and then using LDA to discover the topics on this "document."

Being able to semantically identify entities in content requires that we can disambiguate the entities within the content. In fact, disambiguating entities is a (relatively) old problem in natural language processing (Collins and Singer 1999) and there has been previous work on using dictionaries to aid this task (e.g., (Yarowsky 1992)). We instead leverage Wikipedia as a knowledge base. Other research has proceeded in this direction, leveraging Wikipedia as the knowledge base for entity disambiguation (and labeling) (Kulkarni et al. 2009; Mihalcea and Csomai 2007; Milne and Witten 2008; Cucerzan 2007). We note a key difference is that none of these approaches are leveraged to determine the topics that a user writes about, but rather are mechanisms for disambiguating entities in text. Also, recent work proposed a framework for micro-blogging that includes the capacity to link entities within the post to their disambiguated concept on the Semantic Web (Passant et al. 2008). However, this approach relies on the poster to manually annotate the entities in the post.

3 Building User Profiles From Content

The primary focus of this paper is to study retweeting behavior by building models surrounding user's topic-of-interest profiles. We here spend some time discussing exactly how we create those profiles as they are key to understanding the behavior models we develop later.

We here adopt an approach recently used to profile users in Twitter (Michelson and Macskassy 2010). While the approach was used specifically to profile Twitter users, it can work with any user-generated content. As our study focuses on Twitter, this approach is a good fit for us as we can leverage this to focus on the behavior models below.

We note the real-world nature of Tweets means they are noisy and complex, making our problem difficult. Tweets are intentionally short (limited to just 140-characters) which forces users to be creative in how they constrain the text while preserving meaning. As with text messages in general, this leads to noise. Users rely on common acronyms (e.g., "d/r" means "dressing room" in sports), disambiguation via context ("Arsenal" in a Tweet is the soccer team and not the car, because other soccer players are mentioned in the Tweet), combinations of the two ("Hawks" means "Chicago Blackhawks," if the Tweet mentions "Chicago"), and other

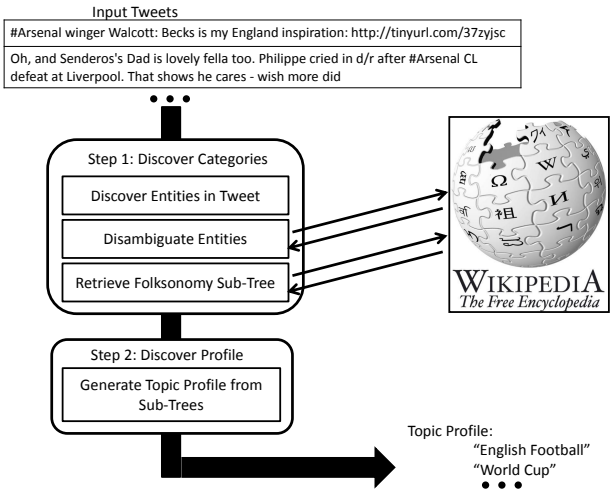

Figure 1: Topic Profiles from User-generated Content

constraining mechanisms. However, Tweets can also be information rich, because users tend to pack substantial meaning into the short space.

The aim is to generate "topic profiles" of users based upon what they post about. We define a topic profile as a list of the common, high-level topics about which a user posts, under the premise that these are the topics of interest to a Twitter user, since s/he Tweets frequently about them.

Our approach to discovering a Twitter user's topic profile is based open the idea that the topics of interests can be identified by finding the *entities* about which a user Tweets, and then determining a common set of *high-level categories* that covers these entities. As a running example, consider the following real-world Tweet:

`#Arsenal winger Walcott: Becks is my England inspiration: http://tinyurl.com/37zyjsc`

There are four entities of interest in this Tweet: Arsenal, which refers to the Arsenal Football Club of England; Walcott, which refers to Theo Walcott, a player for Arsenal; Becks, which refers to football superstar David Beckham; and England. A category that covers these entities within the Tweet might be "English Football." Therefore, to develop a topic profile for a user, we analyze all of their Tweets and determine the set of common high-level categories that covers the set of Tweets. This set of categories defines the topic profile. In our example, the profile may include "English Football," "World Cup," etc.

In order to map entities into high-level topics, and following other prior work in this space, we here use Wikipedia as a knowledge base. Wikipedia provides encyclopedic knowledge about entities which we leverage to disambiguate their mentions in the Tweets. Once disambiguated, we use the "folksonomy"[1] defined by Wikipedia's user-defined categories to map entities to the categories that will define the topic profile.

The general approach consists of two steps and is shown in Figure 1. In step 1 ("Discover Categories"), we discover the entities in the Tweets, disambiguate them, and then retrieve the sub-tree of categories from the folksonomy that contains the disambiguated entity. In step 2 ("Discover Profile"), we analyze all of the subtrees for all of the discovered entities in a users set of Tweets, and determine the set of categories that defines that user's topic profile (e.g., the topics of interest).

[1]A folksonomy is a crowd-sourced taxonomy

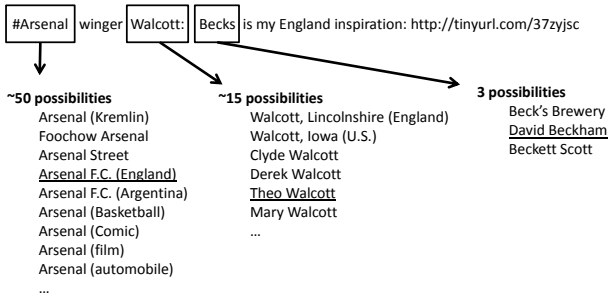

Figure 2: Candidate Entity Matches from Wikipedia

3.1 Discovering Categories for Tweets

The first step in discovering the categories for Tweets involves discovering the entity mentions in the Tweets themselves. As mentioned above, this can be challenging because Tweets are non-grammatical (precluding parsing) and sometimes all capitalized (or all lowercased). Generally, the task of discovering entities is called "named entity recognition" (NER). While much work on NER first parses sentences and finds phrases that include proper nouns, Tweets are ungrammatical and noisy, and we therefore cannot guarantee parses for our data. Our approach is to look for capitalized, non-stopwords as possible named entities. This ensures high recall (we retrieve many possible entities) while conforming to the difficulty of our data. Using our running example Tweet, the entities discovered in this manner are underlined:[2]

`#Arsenal winger Walcott: Becks is my England inspiration: http://tinyurl.com/37zyjsc`

Once we have discovered the entities in a Tweet, we next disambiguate them by finding the page in Wikipedia about that entity. If the entity is not found in Wikipedia then we do not include it in our profile. While we realize that this is apt to miss many entities, our focus in the paper is to identify the higher level ontological concepts within Wikipedia rather than the entity specifically. We note that this is an area of further research we are looking into.

The way we identify the specific Wikipedia page for an entity is to search for the entity (looking for the entity either in the text of a page or in the title). Wikipedia may return a set of candidates that match the entity. For example, Figure 2 shows a subset of pages returned by Wikipedia for the entities identified in the example Tweet, with the correct candidate entity underlined. As the figure shows, for some entities, the disambiguation requires deciding between a large number of possibilities.

To deal with the disambiguation problem, we leverage the "local context" of the Tweet. Specifically, we treat the text of the Tweet (excluding the entity term to disambiguate) as the context for that entity. If we are using the example

[2]Note, we ignore the # sign which is specific for creating within Twitter search links.

Tweet, and our current entity to disambiguate is "Arsenal," then the local context is {winger, Walcott, Becks, ...}. Again, note that we exclude stopwords from the context. More formally, we define the Tweet's local context, C_T, for an entity, E_T, as:

$$C_T \Leftrightarrow \{(t_T \in T_T)/E_T\}$$

where T_T is the set of terms in the Tweet.

We define each candidate entity from Wikipedia as $e_i \in E$ (the set of candidates), and define the context for the page of each candidate entity as:

$$C_{e_i} \Leftrightarrow \{(t_{e_i} \in T_{e_i})/e_i\}$$

We then select the entity from Wikipedia that has the best overlap with the entity in the Tweet, given the local context. In other words, we choose entity e_i from the set of entity candidates E with the maximum contextual overlap:

$$\arg\max_{e_i \in E}(C_T \cap C_{e_i})$$

Remember that we are interested in the higher concepts which are relevant to the entity in question. We retrieve these as a category tree based on the folksonomic category tree which the identified entity page is situated in.

This is done by following the categories which can be found at the bottom of most Wikipedia pages. Each such category has a name, and links to its category page. That category page in turn contains a list of entities that belong to that category, along with another set of categories that generalize the current one (e.g., parent categories).

In our approach we start with the set of categories for the given entity, and trace through the links of each category, collecting the parent categories along the way. At the end of this process, we have a "sub-tree" of the folksonomy, rooted at the most specific term (the current entity's categories). Our sub-trees are actually upside down in that the deeper levels in our trees actually represent more general categories. We discuss below how we handle this characteristic. Also, note, we empirically chose to go 2 levels deep, as at this point the categories were sufficiently general and vague, and with a branching factor averaging around 20, this already provides a large number of categories.

An example of building the sub-tree by walking through the categories of a given entity is shown in Figure 3. In this example, we start with the category "English footballers," from Theo Walcott's page, which produces four categories: {"English Sportspeople," "Association football Players by nationality," "Football in England," and "British footballers"}. Then we show tracing this one step further, by following the categories from "Association football Players by nationality." The resulting sub-tree is also shown.

As in previous work (Ponzetto and Strube 2007)[3] we ignore the categories related to the administration of Wikipedia itself, such as categories with names "Category," "Wikipedia," "Template," etc.

3.2 Generating User Profiles from Category Trees

The output of the previous step is a set of sub-trees rooted on the categories for each of the disambiguated entities in each

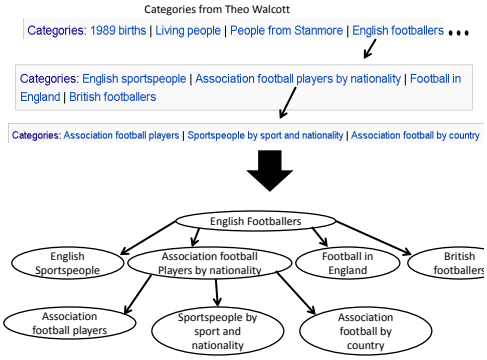

Figure 3: Part of the Sub-Tree from an entity

of the Tweets. The goal of this step is to take in this forest of sub-trees and generate a user-specific topic-of-interest profile. Because the categories are generated from the disambiguated entities, we can assume that the categories (nodes in the subtrees) already cover the entities in the Tweets. We can naïvely generate a user profile by tallying up the frequency of all the category-nodes which we have seen for a given user. However, as noted above, the root categories are much more specific than the other categories (as they are specifically about the entities) and the more general categories tend to show up much more frequently. To even the counts, we weight categories by their depth in the tree and then rank each of the categories, c, in the set of sub-trees according to the following ranking function:

$$Rank(c) = Freq(c) * w_c,$$

where $Freq(c)$ is the frequency of the category's occurrence and w_c is a weight, inverse to the category's level in the sub-tree. Assuming branching factor b, and depth d in the tree for our current category, we define w_c as:[4]

$$w_c = 1/b^d$$

Because our knowledge base is a folksonomy, it may be inconsistent, having categories that occur at various depths, in different parts of the ontology. Therefore, we note that the ranking score of a category is actually the sum of its ranking scores for each depth where it occurs. For instance, if a category occurs 4 times in the 2nd level and 8 times in the 3rd level, its Rank would actually be $4 * (1/b^2) + 8 * (1/b^3)$.

Finally, we define a user's topic profile as the complete set of all observed categories for that user, ranked according to our ranking function.

4 Information-propagation Behavior Models

The primary thesis which we explore in this paper is that people are more likely to propagate information which they find interesting and worth sharing than random content which they see. While it may be difficult to always know what a person finds interesting, it stands to reason that if a person posts about certain topics, then that person is interested in those topics.

[3]The authors focus on building a taxonomy from Wikipedia

[4]Note that we empirically set b to be 10.

If we can encapsulate this information, then we can directly use it to build a more targeted information-propagation model which can better explain the information-propagation behavior we observe in social media. One of the keys to our approach is that we are interested in understanding retweeting behaviors at the individual level rather than at a global level. Therefore, we are not interesting in modeling the message by itself but rather how the message resonates with the individuals who may decide to pass it on.

We therefore define the problem of information-propagation as one of having an individual decide whether to retweet or propagate an observed Tweet. As a user processes a stream of Tweets, at some point a decision is made to retweet one of the observed Tweets. We want to explore different models for selecting which Tweet that would be. In other words, we want to compute $P(retweet(x))$, where x is a Tweet previously seen (up to and including the most recent Tweet).

We next define our four retweeting models.

4.1 General Model (general)

The most basic model assumes that a user will randomly retweet any Tweet previously seen, but with a much higher likelihood of retweeting a Tweet just seen than one seen longer ago. As we will see later, if we map how long ago a Tweet was originally posted before it was retweeted, we get a distribution which looks very much a powerlaw. We therefore use a powerlaw to represent our general retweeting model:

$$P_{\text{gm}}(x) = \alpha * time(x)^{-\beta},$$

where $P_{\text{gm}}(x)$ is the likelihood that x will be retweeted and $time(x)$ is defined as the number of minutes passed since x was original tweeted. We will fit this model to the data below.

4.2 Recent Communication Model (recent)

Second, we consider the network and recency effect, where a user may be more likely to retweet someone s/he has recently been in "contact" with either through a retweet or through a direct message (by using the `@user` construct in Twitter, which can be interpreted as a direct public communication to the given user).

We modify the general propagation model to especially consider Tweets by someone the user has recently been in contact with. This model is defined as:

$$P_{\text{recent}}(x) = P_{\text{gm}}(x) * \left[\begin{array}{l} \alpha * P(x|I(x)) + \\ (1-\alpha) * P(x|!I(x)) \end{array} \right],$$

where α is a parameter we estimate for the likelihood that a user will retweet someone which the user has been in "contact" with within the last 24 hours, and $I(x)$ (and $!I(x)$) represents the fact that x was Tweeted by someone in the set (and *not* in the set) of recent "contacts".

4.3 On-topic Model (topic)

We now turn to our content-based models. The first model we explore is whether a person is more likely to retweet a Tweet which is aligned with the user's topic-of-interest profile. Remember that a user's profile is set of weighted Wikipedia categories and a Tweet is also a set of weighted Wikipedia categories. We represent these sets as vectors in high-dimensional space, where each dimension represents one Wikipedia category.

We define the *similarity* between a Tweet and a user's profile as the *cosine distance* of the two category vectors. By observing what is retweeted, we can generate the underlying empirical distribution of $P_{\text{ts}}(x|sim_T(x,u))$, where $sim_T(x,u)$ is the similarity between a user's profile and that of the Tweet. Our topic-based model is then defined as:

$$P_{\text{topic}}(x) = P_{\text{gm}}(x) * P_{\text{ts}}(x|sim_T(x,u)).$$

Because our empirical model $sim_T(x,u)$ comes from the data, we may find that there are certain levels of similarity where a user is more likely to retweet. For example, we may see some extreme behavior where someone retweets only Tweets which are aligned with the user's own posts, or someone who only retweets Tweets which are far from the persons own interests. The former behavior would suggest very topic-specific information diffusion because people only retweet things of their own interest. The latter suggests a more diverse propagation paradigm, where users tend to only pass on "surprising" information outside their own topic of interests.

4.4 Homophily Model (profile)

The final retweet model we consider here is based on profiles of users. It may be that a user is more likely to retweet another user if they share similar profiles. In other words, is they are similar in their interest, perhaps they are more likely to retweet each other because they find each other's Tweet's interesting even if they are not aligned well with their general interest they post on.

We define similarity as above and represent the profiles as vectors in high-dimensional space as described above.

By observing what is retweeted, we can generate the underlying empirical distribution of $P_{\text{ps}}(x|sim_P(x,u))$, where $sim_P(x,u)$ is the similarity between a user's profile and that of the profile of the user who sent the original Tweet. Our Profile-based model is then defined as:

$$P_{\text{profile}}(x) = P_{\text{gm}}(x) * P_{\text{ps}}(x|sim_P(x,u)).$$

As above, $sim_P(x,u)$ is an empirical model which comes from the data, and we may find that there are certain levels of similarity where a user is more likely to retweet. As above, there are two interesting extreme behaviors which might come out of this.

5 Case Study

We now turn to our case study. We focus here on a Twitter data set which we have processed using the approach above to generate user profiles. We explore which of our four models best fit the observed retweeting behaviors in the data.

5.1 Data

We are continuously collecting Tweets from a set of about 30,000 Twitterers in the Middle East to explore a geographically constrained set of individuals. We identified these individuals using a snowball sampling method where We started from a seed set of ˜125 Twitterers who self-reported (in their profile) to reside in the Middle East. From there,

we expanded the set of users to monitor whenever we saw a retweet or a mention (the `user` construct), adding only users who self-reportedly were in the same region. After a short period of time, we had reached ˜30,000 Twitterers which turned out to be a fairly stable set of users and we have kept this set since then. The snowball sampling has yielded a constrained set of users who make up on large connected component. It is not unreasonable to assume that we have a slice of Tweets that many of them are aware of, or at the very least is representative of a geographically focused set of Tweets they are likely to see. Clearly this is a geographically biased sample, but it is also powerful because it is thusly constrained and is therefore quite useful as a deeper study in a geographic region.

We have been monitoring these Twitterers since early September 2010, and have tagged the first month of data (9/20/2010 through 10/20/2010), which makes up our data set. The full tweet data set for this period consists of $768,000$ Tweets.

From here, we first down-selected to users who had at least one tweet where we could identify at least one entity (this is not counting retweets) to ensure a minimally valid profile. This resulted in $11,423$ users being selected and $353,690$ Tweets being included. Of these, $115,943$ Tweets were retweets (roughly 32%). We use this data to fit our models below.

For our retweet study we further down-selected to only consider users who retweeted at least five times. This down-selection resulted in us having 1582 users and 79917 retweets.

5.2 Experimental Methodology

The purpose of our study is to explore which of our information propagation models are best are explaining the retweet behavior we see. We answer this question by going through the $79K$ retweets and computing for each of our models how likely that particular Tweet was to be retweeted.

We identify the most likely model for each retweet by calculating the respective probabilities (e.g., $P_{\text{gm}}(x)$) and choosing the model with the highest probability. We sum up over all retweets how often each model was the most likely.

We also compute, for each user, the overall best model for that particular user. We do so by computing, for each model, the overall likelihood of seeing all retweets for a given user. We then compute, for each model, how many users were best explained by that model.

Finally, we compute for each user a profile of how often each model best explained each of the user's retweets. This way we can get a behavior profile of a given user to see if the user consistently is best explained by a specific model.

5.3 Fitting the Models

We first fit our models to the data. While there may be statistical issues with fitting and evaluating the models on the same data if one wanted to use the models for predictive behavior, we are here particularly interested in which models best fit (and hence explain) the data.

We first fit the general model by fitting it to the general distribution of the minutes between a retweet and the original tweet. This distribution seem to follow a powerlaw distribution as we see in Figure 4 and when we fit our general

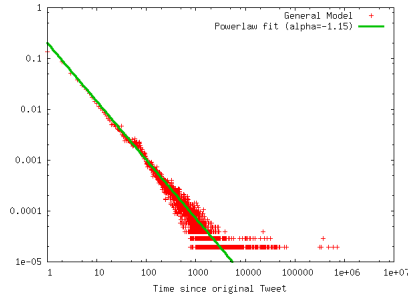

Figure 4: General Model: y-axis is the ratio of retweets, and the x-axis is the number of minutes between a retweet and the original tweet. As can be seen, this approximates a powerlaw distribution with a slope of -1.15.

model to this distribution, we get the following powerlaw distribution:

$$P_{\text{gm}}(x) = 0.2 * time(x)^{-1.15}.$$

We next fit the "recency" model. When we consider all retweets, roughly 38% of them were a retweet of someone who the Twitterer had communicated with (through a retweet or a mention) within the last 24 hours. Hence, the recency model is instantiated as:

$$P_{\text{recent}}(x) = P_{\text{gm}}(x) * \left[\begin{array}{l} 0.38 * P(x|I(x)) + \\ 0.62 * P(x|!I(x)) \end{array} \right].$$

To fit the Topic model and Profile mode, we need to generate the distributions to compute $P_{\text{ts}}(x|sim_T(x,u))$ and $P_{\text{ts}}(x|sim_P(x,u))$. To do this, we computed, for each retweet, the similarity between the Tweet categories and the user profile (of the user doing the retweet) as well as the similarity between the profiles of the user doing the retweeting and the user posting the original Tweet. Our similarity measure (the cosine distance) lies in the range $[0:1]$, which we discretize into 101 bins (0 through 100). We then computed, for each bin, the ratio of retweets which fell into that bin (separately for the profile-similarity and the topic-similarity). These ratios then make up the empirical distribution which is the fit for our two models. To get a sense for whether these fits actually contained any signal, we also computed the "expected" ratios. This was done by computing the average similarity between any tweet and a user (for the expected fit of the topic-model with no observed retweeting behavior) as well as the average similarity between users (for the expected fit of the profile-model again with no observed retweeting behavior).

Figure 5 shows the empirical distributions for the topic-model and the profile-model. Note that the y-axis is a log-scale. We show, for both distributions, the observed retweet fit and the expected fit. Although we can see that all distributions have a very high probability mass at 0 similarity, we see that the expected models had much higher masses at 0 and much lower probability mass as the similarity increased, whereas the observed models showed that there was a strong signal in the similarity. We titled the paper that anti-homophily wins because a 0 mass is still the general winner, although clearly similarity plays a significant role in retweeting behavior.

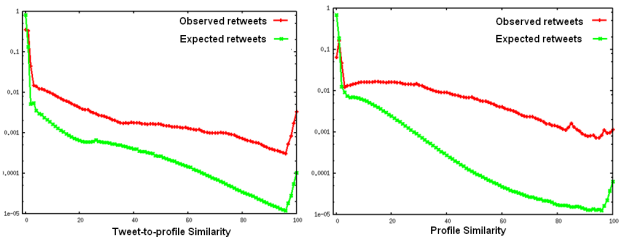

Figure 5: Empirical Distributions of how well the topic of a retweet matches the profile of the person retweeting (on the left), and how well the profile of the original twitterer of a retweet matches the profile of the person doing the retweeting (figure on the right). Note the log scale on the y-axis. See text for details.

Model	General	Recent	Topic	Profile
	10.2%	16.0%	51.6%	22.2%

Table 1: How often was each model the most likely explanation for a retweet.

The question we will answer below in the study is whether the content-based models in fact are an overall better fit when we consider the specific retweets, or whether these observations on the data are just general behaviors which do not translate to individual retweeting behaviors.

5.4 Results

We first explore how often each model best explained each of the $79K$ retweets in our data set. Table 1 shows, for each of the four models, how often it was the best explanation for an observed retweet.

As we can see, the profile-model was by far the most likely model across the board (covering 51.6% of the retweets), followed by the recency model, the topic-based model and finally the general model. While the profile-model weighs anti-homophily highly, the lift for having any similarity clearly shows that people have a strong tendency to retweet others who are like time moreso that topics which are aligned with their own profile.

It may be that some losses are insignificant while wins are significant. It may also be that certain behaviors are more prominent with high-volume users, disproportionally making it show up. To address this, we next explore whether this per-retweet behavior follows into general behaviors by users. Table 2 shows how often each model was deemed the "best fit" for a particular user's specific retweet. As we can see, the qualitative behavior is roughly the same although the general model drops significantly (from 10.2% to 3.9%) and the profile-based model drop slightly (from 51.6% to 47.4%). Most of the gain is in the topic-based model, bringing it almost to a tie with the recency-model. As we can see, the two content-based models account for almost 75% of the retweeting.

Finally, we wanted to explore if the above result shows a consistent behavior of users (i.e., that they tend to follow a single model), or if their behavior really is more spread out among the different propagation models. Table 3 shows how often models were used, on average, across all users. Interestingly, on average the models are used in a very similar distribution to that seen in Table 2, with a 7% drop in the profile-based model which is picked up by the general model. Clearly there is high variation within users from this average profile. For example, for all models it was the case that there was at least one user who never used that model as well as at least one user who exclusively fit that model.

Model	General	Recent	Topic	Profile
Number	61	453	420	750
Ratio	3.9%	28.6%	26.5%	47.4%

Table 2: How many users were best "explained" by each model as the most likely explanation for that user's overall behavior. As we can see, the profile-based model is the better model a significant amount of the time, although other models are also used a significant amount of the time.

Model	General	Recent	Topic	Profile
average	11%	26%	26%	37%

Table 3: How often was each model used by each user.

Number of models used			
One	**Two**	**Three**	**Four**
67	368	568	579

Table 4: How many user behaviors were best explained by a combination of one, two, three or all for models.

We explored the variance across users in more detail to understand just how varied a user's behavior was. Table 4 shows, for each user how many of the four models were the "best fit" at least once among all the retweets of that user. The table clearly shows that over 70% of the users were explained by three or four models, supporting our intuition that users tend to use widely different reasons when deciding whether to retweet.

All of these results indicate strongly that multiple models ought be used and explored when trying to understand why a particular information propagation or information diffusion pattern appears. We have shown that there is a significant effect by the content but that clearly multiple processes are participating.

6 Discussion

We have in this paper taken a close look at what drives certain information diffusion processes in social media. In particular, we studied a set of Twitter users over a period of a month and sought to explain the individual information diffusion behaviors, as represented by retweets, in this domain.

We hypothesized that knowing more about the user and the content would allow us to develop richer models which would take profiles and tagging into account. Specifically, we took an approach to tag Tweets with Wikipedia categories and aggregate these tags for a particular user to generate a topics-of-interest profile for that user. We used these profiles to model retweeting behaviors based on similarities between users and users and the tweets they retweeted.

We explored four retweeting models, two of which were based on user profiles. We found that indeed the content-based propagation models were better at explaining the majority retweet behaviors we saw in our data. When digging

deeper, however, we found that all four models were used at different times and indeed the majority of user retweeting behaviors was best explained by multiple models.

This work is a first step into exploring how to leverage content to generate profiles and context in social media, in order to get a deeper understanding of what drives people to propagate or diffuse information. Specifically, we focused on modeling individual micro-cosm behavior rather than general macro-level processes.

References

Borau, K.; Ullrich, C.; Feng, J.; and Shen, R. 2009. Microblogging for language learning: Using twitter to train communicative and cultural competence. In *Proceedings of the International Conference on Advances in Web Based Learning*.

Chen, J.; Nairn, R.; Nelson, L.; Bernstein, M.; and Chi, E. 2010. Short and tweet: experiments on recommending content from information streams. In *Proc. of the international conference on Human factors in computing systems*.

Clauset, A.; Newman, M. E. J.; and Moore, C. 2004. Finding community structure in very large networks. *Physical Review E* 70. 066111.

Collins, M., and Singer, Y. 1999. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Cucerzan, S. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Procedings of EMNLP-CoNLL*.

Gomez-Rodriguez, M.; Leskovec, J.; and Krause, A. 2010. Inferring networks of diffusion and influence. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Krishnamurthy, B.; Gill, P.; and Arlitt, M. 2008. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks*.

Kulkarni, S.; Singh, A.; Ramakrishnan, G.; and Chakrabarti, S. 2009. Collective annotation of wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *In* Proc. of the International Conference on World wide web.

Lerman, K., and Ghosh, R. 2010. Information contagion: an empirical study of spread of news on digg and twitter social networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*.

Lerman, K., and Hogg, T. 2010. Using a model of social dynamics to predict popularity of news. In *Proceedings of 19th International World Wide Web Conference (WWW)*.

Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N.; and Hurst, M. 2007. Cascading behavior in large blog graphs. In *SIAM International Conference on Data Mining (SDM)*.

Leskovec, J.; Lang, K.; Dasgupta, A.; and Mahoney, M. 2008. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. arXiv:0810.1355v1.

Michelson, M., and Macskassy, S. A. 2010. Discovering users' topics of interest on twitter: A first look. In *Proceedings of the Workshop on Analytics for Noisy, Unstructured Text Data (AND)*.

Mihalcea, R., and Csomai, A. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*.

Milne, D., and Witten, I. H. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*.

Muff, S.; Rao, F.; and Caflisch, A. 2005. Local modularity measure for network clusterizations. *Physical Review E* 72(056107).

Newman, M. 2005. Modularity and community structure in networks. In *Procedings of the National Academy of Sciences*, 8577–8582.

Passant, A.; Hastrup, T.; Bojars, U.; and Breslin, J. 2008. Microblogging: A semantic and distributed approach. In *Proceedings of Workshop on Scripting for the Semantic Web*.

Ponzetto, S. P., and Strube, M. 2007. Deriving a large scale taxonomy from wikipedia. In *Proceedings of the 22nd national conference on Artificial intelligence (AAAI)*.

Porter, M. A.; Onnela, J.-P.; and Mucha, P. J. 2009. Communities in networks. *Notices of the AMS* 56(9):1082–1097, 1164–1166.

Sadikov, E.; Medina, M.; Leskovec, J.; and Garcia-Molina, H. 2011. Correcting for missing data in information cascades. In *ACM International Conference on Web Search and Data Minig (WSDM)*.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the International Conference on World Wide Web*.

Sankaranarayanan, J.; Samet, H.; Teitler, B. E.; Lieberman, M. D.; and Sperling, J. 2009. Twitterstand: news in tweets. In *Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*.

Shamma, D. A.; Kennedy, L.; and Churchill, E. F. 2009. Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media*.

Suh, G.; Hong, L.; Pirolli, P.; and Chi, E. H. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Second IEEE International Conference on Social Computing (SocialCom)*, 177–184.

Weng, J.; Lim, E.-P.; Jiang, J.; and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*.

White, S., and Smyth, P. 2005. A spectral clustering approach to finding communities in graph. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*.

Yarowsky, D. 1992. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of the 14th conference on Computational linguistics*.