City Research Online

City, University of London Institutional Repository

Citation: Bastos, M. T., Travitzki, R. & Puschmann, C. (2012). What Sticks with Whom? Twitter Follower-Followee Networks and News Classification. Paper presented at the The Sixth International AAAI Conference on Weblogs and Social Media (ICWSM 12), 4 Jun 2012, Dublin, Ireland.

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/15366/

Link to published version:

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

What Sticks With Whom?
Twitter Follower-Followee Networks and News Classification

Marco Toledo Bastos
Faculty of Communication and Arts
University of São Paulo
Cidade Universitária, 05508900
São Paulo, Brazil
opus@usp.br

Rodrigo Travitzki
Faculty of Education
University of São Paulo
Cidade Universitária, 05508900 São
Paulo, Brazil
travitzki@usp.br

Cornelius Puschmann
Department of English
University of Düsseldorf
Universitätsstraße 1, 40225
Düsseldorf, Germany
cornelius.puschmann@uni-duesseldorf.de

Abstract

In this paper we analyze Twitter as a news channel in which the network of followers and followees significantly corresponds with the message content. We classified our data into twelve topics analogous to traditional newspaper sections and investigated whether the spread of information depended upon the Twitter network of followers and followees. To test this, we mapped the social network related to each topic and calculated the occurrence of retweet and mention messages whose senders and receivers were interconnected as followers and followees. We found that on average 10% of retweets (RT-messages) and 5% of direct mentions between users (AT-messages) in Twitter hashtags are sent and received by users interconnected as followers and followees. These figures vary considerably from topic to topic, ranging from 15%-19% within Technology, Special Events and Politics to 3%-5% within the categories Personalities and Twitter-Idioms. The results show that hard-news messages are retweeted by a considerably larger community of users interconnected as followers and followees. We then performed a statistical correlation analysis of the dataset to validate the classification of hashtag in news sections based on retweet connectivity.

1. Twitter as a Source of News

Recent literature has examined a number of approaches to information diffusion in Twitter. Previous studies (Bakshy, Hofman, Mason, & Watts, 2011; Huberman, Romero, & Wu, 2009; Kwak, Lee, Park, & Moon, 2010) have shown that Twitter's topological features comprise a highly skewed distribution of followers and low rate of reciprocated ties. Influence on Twitter was found to be connected to network topology, even though metrics such as the number of followers, page-rank, and number of retweets presented different results (Kwak, et al., 2010; Wu, Hofman, Mason, & Watts, 2011).

Bakshy et al. (2011) investigated the distribution of retweet cascades on Twitter and determined that although users with large follower counts and past success in triggering cascades were on average more likely to trigger large cascades in the future, these features were in general poor predictors of future cascade size. Wu et al. (2011, p. 3) found that Twitter does not conform to the usual characteristics of social networks, which exhibit much higher reciprocity and far less-skewed degree distributions, but instead better resembles a mixture of mass communication and face-to-face communication.

Kwak et al. (2010) crawled the entire Twitter network and found a non-power-law follower distribution, a short effective diameter, and low reciprocity, which all mark a deviation from the characteristics of human social networks described by Newman (2003). Kwak et al. also found that Twitter and Korean social network Cyworld present a much higher power-law distribution than most social networks. The characteristics shared by Twitter and Cyworld are that many celebrities are present and that they interact with their fan base.

This characteristic emphasizes the importance of celebrities and media-pundit users in social networks such as Twitter. Kwak et al. (2010) encountered a short average path length that might be a symptom of Twitter's role as an information mechanism, as users follow users not for social networking, but for information. The investigation of Wu et al. (2011) was consistent with the results of Kwak et al. (2010) regarding the topological features of Twitter followers graph. They concluded from the highly skewed nature of the distribution of followers and the low rate of reciprocated ties that Twitter more closely resembled an information sharing network than a social network.

The question of whether Twitter better resembles an information sharing network or a social network was also addressed by exploring the variety of topics that flow throughout the Twitter network. Romero et al. (2011) examined the hypothesis that hashtags for different topics spread differently. The researchers classified Twitter hashtags from a large dataset into eight different topics: Political, Idioms, Celebrity, Sports, Music, Technology, Movies/TV and Games. They concluded that there is significant variation in the mechanics of information diffusion in relation to topics. Following this latter approach, we examined whether Twitter content can be organized like news sections and therefore subjected to principles of newsworthiness.

1. Newsworthiness Criteria

The key factors governing the newsworthiness of information were originally defined by Otto Groth (1928) and included seven newspaper qualifications and a number of article attributes, including relevance, universality, publicity and periodicity. Galtung and Ruge (1965) further explored these categories and identified thirteen factors tested against the hypotheses of additivity, complementarity and exclusion. These principles could then be used to predict how likely it was that a certain event was to be judged newsworthy.

Galtung and Ruge's original research featured a dataset extracted from three major international crises. The data used for the analysis was therefore hard news and did not include soft news articles (see section 4 below). Tunstall (1971, p. 21) commented that because Galtung and Ruge's dataset was restricted to the coverage of international crises, they ignored day-to-day coverage of lesser, domestic and mundane news. This led to further research on the factors driving newsworthiness and to a general consensus that the context of print media is one of an increasing editorial emphasis on entertainment (Franklin, 1997, p. 72).

Harcup and O'Neill (2001) commented on Franklin's work and pointed out that no contemporary set of news values can be complete without the entertainment factor. The authors offered a revised version of Galtung and Ruge's original set of factors, which is similar to the original but includes the factors of Entertainment and Good News (in opposition to Bad News) and the merging of factors Consonance, Composition and Unambiguity into what Harcup and O'Neill called Newspaper Agenda.

Galtung and Ruge's News Factors (1965)	Harcup and O'Neill News Factors (2001)
Personification	The Power Elite
People	Celebrity
	Entertainment
Unexpectedness	Surprise
Negativization	Bad News
	Good News
Threshold	Magnitude
Meaningfulness	Relevance
Nations	
Continuity	Follow-Up
Consonance	Newspaper Agenda
Composition	
Unambiguity	
Personification	
Frequency	

Table 1 Galtung and Ruge versus Harcup and O'Neill news factors

Harcup and O'Neill's (2001) suggestion highlighted the shifting paradigm of newsworthiness from mid-twentieth-century news reporting, focused on political and socio-economic issues, to infotainment news covering celebrities' personal lives and showbiz events. The dominance of celebrity and social news, and the increasing growth of reality shows and other forms of popular-culture oriented news contributed to the blurring of credibility boundaries that once set traditional outlets apart from digital media (Johnson & Kayer, 2004).

Michael Schudson enumerated the decisive factors in the overall change in the news ecosystem, which stems from the collaboration between reader and writer: the lack of ultimate distinctions among tweets, blog posts, newspaper stories, magazine articles or books, and the diminishing gap between professionals and amateurs (Schudson, 2011, pp. 207-216). Schudson's conclusion is that the line between old media and new media has been blurred beyond recognition and that the very nature of news values is evolving. Even though Twitter is experiencing exponential growth in infotainment news, or perhaps precisely because of that, it offers a privileged view of the dynamics of digital news.

2. News Propagation on Twitter

The investigation of Romero et al. (2011) found that the variation between topics was not only a result of stickiness, that is, the probability of adoption based on one or more exposures to the hashtag. The results also indicated a significant difference in the persistence of the hashtags according to the topic in which they were classified. Hashtags with high persistence tended to continue having relatively significant effects even after repeated exposures. Perhaps not surprisingly, Romero et al. (2011) found that hashtags for politically controversial topics were particularly persistent.

The opposite effect was found in the class of hashtags the researchers named Twitter-Idioms, which refer to a type of hashtag familiar to Twitter users in which common words are concatenated into neologisms that serve as a marker for conversational themes. The investigation found that stickiness in Twitter-Idioms hashtags was high, but the persistence was unusually low, meaning that the chance of a user adopting the hashtag fell quickly if the hashtag was not adopted after a small number of exposures.

Romero et al. (2011) stressed that the distinctive network structure of Twitter political hashtags—the unusually large effect relative to the peak after successive exposures—not only corresponded with the sociological principle of complex contagion, but also depicted the first large-scale validation of the principle. On the one hand, Political and Games hashtags emerged as persistent topics because users refer to these keywords many times. Hashtags associated with Twitter-Idioms and Technology, on the other hand, were used by a higher number of users in comparison to other topics, but users tended to use the hashtags only once or a few times, thus rendering a lower number of total mentions in comparison to other topics.

3. News Categories and Twitter Topics

Classifying news is a common problem in professional journalism, where news is purportedly divided between hard and soft news. While hard news coverage relies on fact-checking and research, soft news is often directed by marketing departments and heavily influenced by demographic appeal and audience share. Hard news embodies the principles of seriousness and is based upon a timeline in which the story unfolds. The definition of soft news falls somewhere in between information and entertainment—a conceptual nexus expressed in the neologism "infotainment."

Hard news topics include politics, economics, crime and disasters, but can also encompass aspects of law, science, and technology. Soft news topics include the arts, entertainment, sports, lifestyles, and celebrities. Unlike hard news, soft news stories do not depend upon a timely report, as there is no precipitating event triggering the story other than the public's or the reporter's curiosity. We expected the division between hard and soft news not only to be valid for Twitter topics, but also to be noticeably clear in view of the increasing prominence of infotainment-oriented content (Bourdieu, 1998; Franklin, 1997).

In the following table we gathered the regular sections of a newspaper, classified according to the principles of newsworthiness, together with the topics investigated by Romero et al. (2011) and the topics investigated in this paper. Twitter topics analyzed by Romero et al. (2011) are presented on the extreme left, followed by a general classification of newspapers sections and the topics investigated in this paper.

News	Romero et al. (2011)	Newspaper Sections	Twitter Topics
Hard News	Political	Politics	Politics
		World	
		National	
		Economy	
		Local news	Altruism
			Events
		Science	
↑ ↓	Technology	Technology	Technology
	Games		Games
	Idiom	Opinion	Idioms
Soft News	Music Movies TV Celebrity	Arts Entertainment Environment Medicine	Music Personality Movies Celebrity
		Fashion	Lifestyle
	Sports	Sports	Sports

Table 2 Similarities between news sections in print media and in Twitter topics

To test our classification we compiled a dataset of approximately 2 million messages and divided them into the 12 aforementioned categories. The proposed classification reflects the increasing trend towards entertainment news. While traditional newspapers devote one to two sections to soft news, Twitter messages are substantially devoted to it, as messages related to Music, Games, Personalities, Movies and Celebrities are responsible for a significant share of Twitter's information stream. The reverse pattern is found in politics, which is covered in multiple newspaper sections but is matched by a single Twitter topic. We matched Twitter-Idioms with the Opinion page in newspapers, as Idioms serve as a marker for a conversational theme while also offering a platform for airing one's opinion.

4. Dataset

We examined Twitter as a news provider using a dataset of 108 hashtags divided into 12 topics, so that each topic consists of 9 hashtags. The dataset spans from 9 February to 28 November 2011, with two-thirds of the hashtags having featured in Twitter Trending Topics. The selection was based on the size of the hashtags, having on average 20,000 tweets each. Immediately after the archiving process we mined the social data for each keyword or hashtag. The topics were categorized as follows: Events, Technology, Politics, Altruism, Games, Lifestyle, Movies, Sports, Celebrity, Music, Personality, and Idioms.

The dataset contains 1,905,989 tweets and over 14 billion non-unique Twitter users, of which 1,017,046 are interconnected as followers and followees. From the nearly 2 million tweets in the dataset, 460,960 are retweets and 42,520 are retweets sent and received by users connected as followers and followees. The total number of AT-messages is 108,261 and a total of 4,892 of these messages were sent and received by users connected among themselves as followers and followees.

Category	Definition
Events	Includes names of days in a concatenation similar to Idioms-hashtags, including public holidays, special days, and historical anniversaries. Event-hashtags refer to a precise date.
Movies	Includes names of film releases and events related to a particular film production. Includes keywords related to original and international releases. It does not include actors or performers who might have worked in the film.
Technology	Includes names of websites, applications, devices, internet services, operational systems, software, computer platforms, and manufacturers. It also includes events specifically involving any of these.
Sports	Includes names of sports teams, matches, leagues, athletes or particular sports events. It also includes references to news items that specifically address sports.
Politics	Includes names and keywords that refer to political events, demonstrations, riots, coups d'états, and marches or simply to a politically controversial topic. It includes political figures, commentators, movements, and parties.
Celebrity	Includes names of persons or groups prominently featured in entertainment news. It does not include politicians, media-pundits or religious representatives. The name of the celebrity is sometimes found in a longer hashtag referring to some event related to the celebrity.
Altruism	Includes the names of events, campaigns, and assemblies aimed at altruistic actions for a local or broader community. It can also contain political declarations and human rights related campaigns.
Music	Includes names of songs, albums, groups, musicians and performers who work with music. It also includes events involving the artists and pools about songs and artists.
Games	Includes names of game vendors, game platforms, game consoles, MMORPG, or twitter-based games, as well as groups devoted to such games.
Personality	Includes names of media personalities who are not considered celebrities, but are frequently featured in media or who work for media outlets. It includes news anchors, journalists, comedians and sports professionals.
Lifestyle	Includes tags and words associated with the way a group or a person lives. It includes behavior and consumer trends, fashion, habits, and advertising

	campaigns.
Idioms	Includes tags referring to a conversational theme that consists of a concatenation of at least two words. The concatenation usually does not include names of people or places and the full phrase is not a proper noun or a reference to the title of a song, movie or organization.

Table 3 Definition of categories applied to hashtags and keywords

Category	Hashtags
Events	diadodoadordesangue; diadofrevo; diadoreporter; diamundialsemtabaco; heliogracieday; dianacionaldovolei; diadoadvogado; parabensnossasenhora; semanadodoador
Movies	pussinboots; moneyball; towerheist; tintin; swath; sherlock; jedgar; mostra; theskinilivein
Technology	wp7; windowsphone; windows8; icloud; rim; iphone4s; ios; galaxynexus; siri
Sports	tanopasman; corinthians; vasco; ufc139; allblacks; rwcfinal; brasileirao; ufc126; ufc132
Politics	sosnatal; battisti; marchadamaconha; m15; freeiran; abaixodecreto; amandagurgel; ukrevolution; globalcamp
Celebrity	katewinslet; leonardodicaprio; eddiemurphy; jovelinadascruzes; axl; kesha; shakira; demiyouarebeautiful; tomastranstromer
Altruism	vaidoa; doadordesangue; adoteumanimalabandonado; marcoule noalaviolenciamachista; realengo; trabalhoescravo; pedofilianao; aligadavida
Music	coldplay; vivalavida; zecabaleiro; gnr; guns; lennykravitz; myfavoriteartist; 14millionbeliebers; lagumalampertama
Games	crysis2; videogamedeals; zelda; mari0; minecon; halo4; lanoire; frugalgaming; wii
Personality	voltarafinha; evaristocosta; imiteomarcoluque; quedeselegante; calabocagalvao; jimschwartz; jackwilshere; freebruce; claymatthews
Lifestyle	cantadasindie; cervejadeverdade; maconha; odeiorodeio; seeufosserico; escolhiesperaremdeus; amorodeio; estudarvaleapena; undateable
Idioms	1bomprofessormeensinou; biggestlessonlearnedfrom911; illpunchuinthefaceif; myworldmemories; otrosusosparaelblackberry; qndomertiolateardia; terriblenamesforavagina; brazilwaits4bustinjieber; favoritenbamoments

Table 4 List of hashtags and keywords in the dataset

5. Methodology

We investigated 108 different hashtags classified according to their content in the following 12 categories: Politics, Events, Idioms, Celebrity, Personality, Sports, Music, Technology, Movies, Lifestyle, Altruism, and Games. Next we mapped the social network of each hashtag and separated the tweets between users interconnected as followers and friends (FF) and users that were not interconnected. After calculating the overall percentages, we found out that messages between users interconnected as followers and followees is on average of 10% for retweets (RT-connectivity) and of 5% for mentions (AT-connectivity).

We estimated that RT-messages whose senders and receivers were interconnected as followers and followees would rely on Twitter's network to spread the information, while RT-messages without interconnected users should rely on other networks, such as media outlets and peer-to-peer communication. Lastly, we ran a statistical correlation analysis to compare the topic classification with the components of each subset, including AT and RT to users, number of tweets, number of users and the total number of followers and followees.

Even though we found that on average only 10% of retweets were sent and received by interconnected users, these figures vary greatly from topic to topic, being as high as 19% in Events and as low as 3% in Idioms. We understand that hashtags and keywords have diffusion patterns connected to the content of the messages, given that the information they contain is intended for different publics.

	Category	RT-connectivity	AT-connectivity
Hard News	Events	19%	6%
	Technology	16%	5%
	Politics	15%	7%
	Altruism	15%	6%
Soft News	Games	12%	2%
	Lifestyle	12%	7%
	Movies	12%	6%
	Sports	11%	3%
	Celebrity	10%	2%
	Music	9%	6%
	Personality	5%	3%
	Idioms	3%	4%

Table 5 RT and AT-messages between users connected as follower and followee

The highest percentage of retweet-messages between connected users was found in the Altruism-hashtag group, being as high as 44% for the hashtag diadodoadordesangue (Figure 4). The lowest percentage of retweet-messages between interconnected users was found in the Idioms-hashtag group, being as low as 1% for the hashtags favoritenbamoments and otrosusosparaelblackberry.

The differences in RT-connectivity allow the detection of incorrectly categorized hashtags. The classification based on content analysis was not conclusive, as a significant portion of the dataset could be assigned to more than one topic. We proceeded to a categorization based on RT-connectivity and the results show that Twitter content can be classified according to user's connectivity. Keywords and hashtags often contain information pertaining to more than one topic. The hashtag biggestlessonlearnedfrom911 was first placed in the Events-hashtag group. But the retweet and mention realization shows that biggestlessonlearnedfrom911 is actually part of the Idioms-hashtag group, in which hashtags have very low retweet and mention realization.

Hashtag	Category	RT connectivity	AT connectivity
dianacionaldovolei	Events	14%	6%
diadofrevo	Events	23%	4%
diadoreporter	Events	25%	7%
diamundialsemtabaco	Events	16%	5%
biggestlessonlearnedfrom911		**1%**	**3%**
diadoadvogado	Events	12%	9%

Table 6 Event messages between users connected as follower and followee

The hashtag otrosusosparaelblackberry was at first assigned to the Technology group, but its RT-connectivity caused it to be recategorized into the Idioms group. The hashtags biggestlessonlearnedfrom911 and otrosusosparaelblackberry have similar percentages of retweet and mention messages among interconnected users, and both were previously placed in groups with connectivities different than their own.

Hashtag	Category	RT connectivity	AT connectivity
windowsphone	Technology	21%	2%
wp7	Technology	25%	5%
windows8	Technology	19%	7%
otrosusosparaelblackberry		**1%**	**3%**
icloud	Technology	18%	7%
iphone4s	Technology	11%	5%

Table 7 Technology tweets between users connected as follower and followee

RT-connectivity also corrected the classification of hashtag heliogracieday, which celebrated the birthday of deceased Jiu-Jitsu grandmaster Helio Gracie. At first we placed the hashtag in the Sports-hashtag group, but the mechanics of retweet and mention diffusion indicates that the hashtag actually follows the Events-hashtag pattern of information replication.

Hashtag	Type	RT connectivity	AT connectivity
heliogracieday		**14%**	**3%**
ufc132	Sports	2%	1%
ufc126	Sports	7%	1%
vasco	Sports	5%	4%
rwcfinal	Sports	5%	1%

Table 8 Sports messages between users connected as follower and followee

User connectivity also shed light onto the question of whether topic-related hashtags create or foster communities within Twitter. We expect AT-connectivity to be a good predictor of community engagement. However, the results show that AT-messages that include a hashtag have an average rate of 5% of interconnectivity between users, and vary little from topic to topic (Figures 1 and 2).

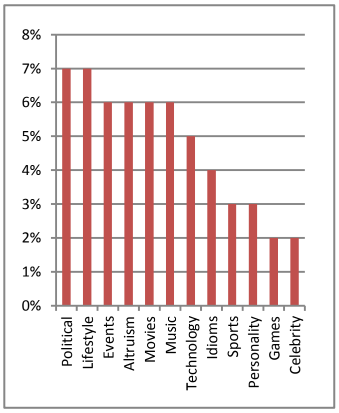

Figure 1 AT-messages among interconnected users classified by topics

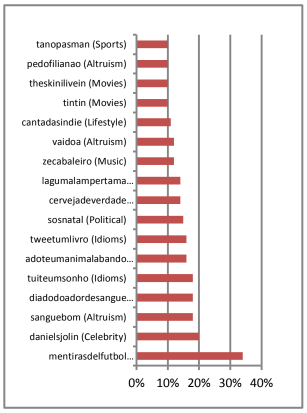

Figure 2 List of top hashtags among interconnected users in AT-messages

Because the variance is significantly low, we concluded that AT-messages are not affected by the interests or the network topology of the users sending and receiving messages. Instead, these messages seem to configure a peer-to-peer conversation that is not related to the broader topic under discussion. Even though some topics presented a variance of 1%-2%, additional analysis based on overall message volume indicates that AT-messages present no significant variation regarding the interconnection of senders and receivers as followers and followees.

6. One Network for Each News Topic

Current studies highlighted that Twitter network structure better resembles an information sharing network than a social network. Nonetheless, our results indicate that Twitter network topology is not of decisive importance to the spread of information, as the network of followers and followees accounts on average for only 10% of message replication. However, the results are consistent with the classification of topics according to hard and soft news, which is a characteristic of media outlets.

Hard-news is retweeted by a considerably larger community of users interconnected as followers and followees, while soft news and Idioms-like hashtags are at the bottom of the rank. Our results can be divided into three groups. At the bottom of the table (Table 5) we find Idioms and Personality hashtags in which messages are retweeted among the smallest percentage of interconnected users (3% and 5%, respectively). These topics cannot be classified as soft-news, as the hashtags and keywords do not focus on arts, entertainment, sports or lifestyles, but instead on a variety of personal statements and infotainment news boosted by the increasing popularity of reality shows and other forms of popular culture. In the intermediary zone we find the actual material of soft-news topics, including Celebrities, Sports, Movies, Lifestyle, and Games (10%, 11%, 12%, 12% and 12%, respectively).

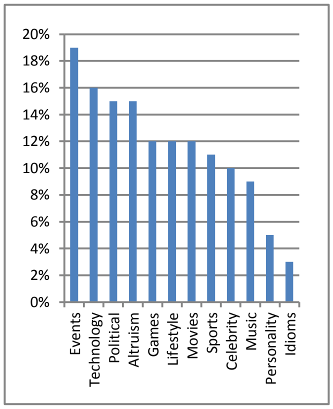

Figure 3 RT-messages among interconnected users classified by topics

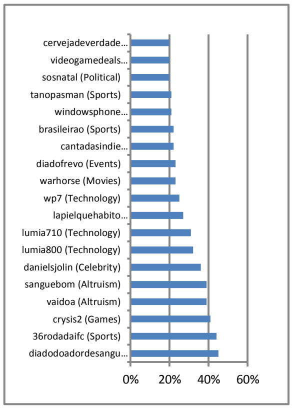

Figure 4 List of top hashtags among interconnected users in RT-messages

On average there are 7% more replicated messages among interconnected users in the soft-news plateau in comparison to the bottom of the table (Figure 3 and Table 5). At the top of the table we found hashtags and keywords related to hard-news topics that correspond to local and national Events, Technology, Politics, and Altruism campaigns (19%, 16%, 15% and 15%, respectively). Again we observe a significantly higher number of retweeted messages sent and received by users interconnected as followers and followees. The difference between soft-news and hard-news is on average 5%.

We found little variation in AT-messages, even though some hashtags presented a considerable difference in the percentage of AT-messages sent among interconnected users (Figure 2). The content of these hashtags are made up of peer-to-peer interaction, including gambling on sports results, gossip about the personal lives of professional athletes, philanthropic campaigns, and Twitter-Idioms hashtags. As we hypothesized at the beginning of the paper, hashtags with a higher incidence of retweet-messages from interconnected users presented significantly higher community-related content and might conform to the characteristics of human social networks.

7. Correlation Analysis of Twitter Topics

We performed a Pearson correlation analysis of the main components of each topic. A correlation coefficient (p<0.001) was computed for each pair of the 17 arrays in the dataset and is presented qualitatively on a matrix, colored in yellow for high correlation and blue for low. High correlations indicate a predictive relationship between units, while low correlations indicate that the arrays do not vary together.

We first looked into the correlations among retweets to users, retweets from users, AT-messages from users, AT-messages to users, and the number of tweets. This set of correlations defines the basic conversational features of each topic, as it correlates RT, AT, and tweets. We found that topics traditionally defined as news magnets, such as Politics, Altruism, Lifestyle, Movies, and Sports have a significantly higher correlation among these five arrays. Topics like Technology, Events, and Celebrity presented correlations only between the number of tweets with AT indegree and RT outdegree, therefore suggesting a relationship among AT-messages, RT-messages, and the number of tweets. Topics like Music, Personality, Idioms and Games presented a lower-than-average set of correlations, being statistically significant only between the number of tweets with AT-messages and with RT-messages.

Next we looked into the correlations among the numbers of tweets, retweets to users and from users, interconnected users, and number of followers. This set of correlations compares the increase of tweets and retweets within and without the Twitter network of followers and followees. We found that topics like Idioms, Personality, Music, Celebrity, Lifestyle, Events, and Technology present significant correlations only between the number of tweets and retweets. Topics like Politics, Altruism, Events, and Games presented correlations between RT-messages and AT-messages within the FF network, thus suggesting that these topics tend to create a conversation within the user's network, not only to perform broadcasting functions.

Altruism-related messages presented a negative correlation between RT to users and user's number of followers, meaning that the more these users retweet Altruism-related messages, the more likely it is that these users have a low number of followers. This result suggests that popular Twitter users engage in Altruism-related messages less often than the average user, possibly because Altruism-messages consist of aiding and rescuing campaigns that are not publicly appealing. We found no correlation between Twitter account creation dates and retweet rate, except in the Technology topic. We regard this result as a reflection of the early adopter profile of users tweeting about technology. We also found a unique correlation match in the Celebrity and Games topics. These were the only topics to present unidirectional correlation between retweets from users and hashtag followers, thus suggesting that followers of a given account tend to retweet the content of that account significantly more than in other topics. We interpret this result to reflect the activity of fan groups surrounding celebrities.

Last we looked into a larger number of correlations to get a sense of which topics mobilize the most elements of Twitter's network (Figure 5). We looked into the correlations among tweet percentage, AT and RT to users, AT and RT from users, AT and RT percentage, AT and RT between connected and non-connected users, number of tweets, number of interconnected users, followers and followees percentage, the number of followers, and the number of tweets of each user's account. Once again we found that topics traditionally defined as news magnets like Politics, Movies, and Lifestyle presented significantly higher total number of statistically significant correlations in comparison to the remaining topics. Infotainment topics like Sports, Celebrity, and Music presented a considerably lower number of correlations among the aforementioned arrays, while Idioms, Personality, Altruism, Games, Technology, and Events lay in between the two groups.

Even though the correlation plots highlight that the division among topics is consistent with news sections, the division itself is at odds with the classification provided by the separation of topics according to retweet interconnectivity of users. This implies that there are other factors driving topic categorization which requires further investigation. However, we found that the mean and median values for Twitter users' account creation date are consistent with the topic division described in section 6.

	Category	**Account Creation Date**				**Category**	**Account Creation Date**		
		Max	**Med**	**Mean**			Max	**Med**	**Mean**
Hard News	Technology	1940	**693**	**671**	**Soft News**	Sports	1809	**525**	**520**
	Altruism	1861	**603**	**570**		Lifestyle	1809	**525**	**520**
	Politics	1916	**599**	**565**		Celebrity	1927	**502**	**512**
	Events	1766	**590**	**551**		Music	1924	**459**	**483**

Table 9 Minimum, Maximum, Median and Mean of Twitter account creation date classified by topics and sorted by highest median. The group on the left has the highest median and the group on the right the lowest. The division is consistent with hard and soft news topic classification.

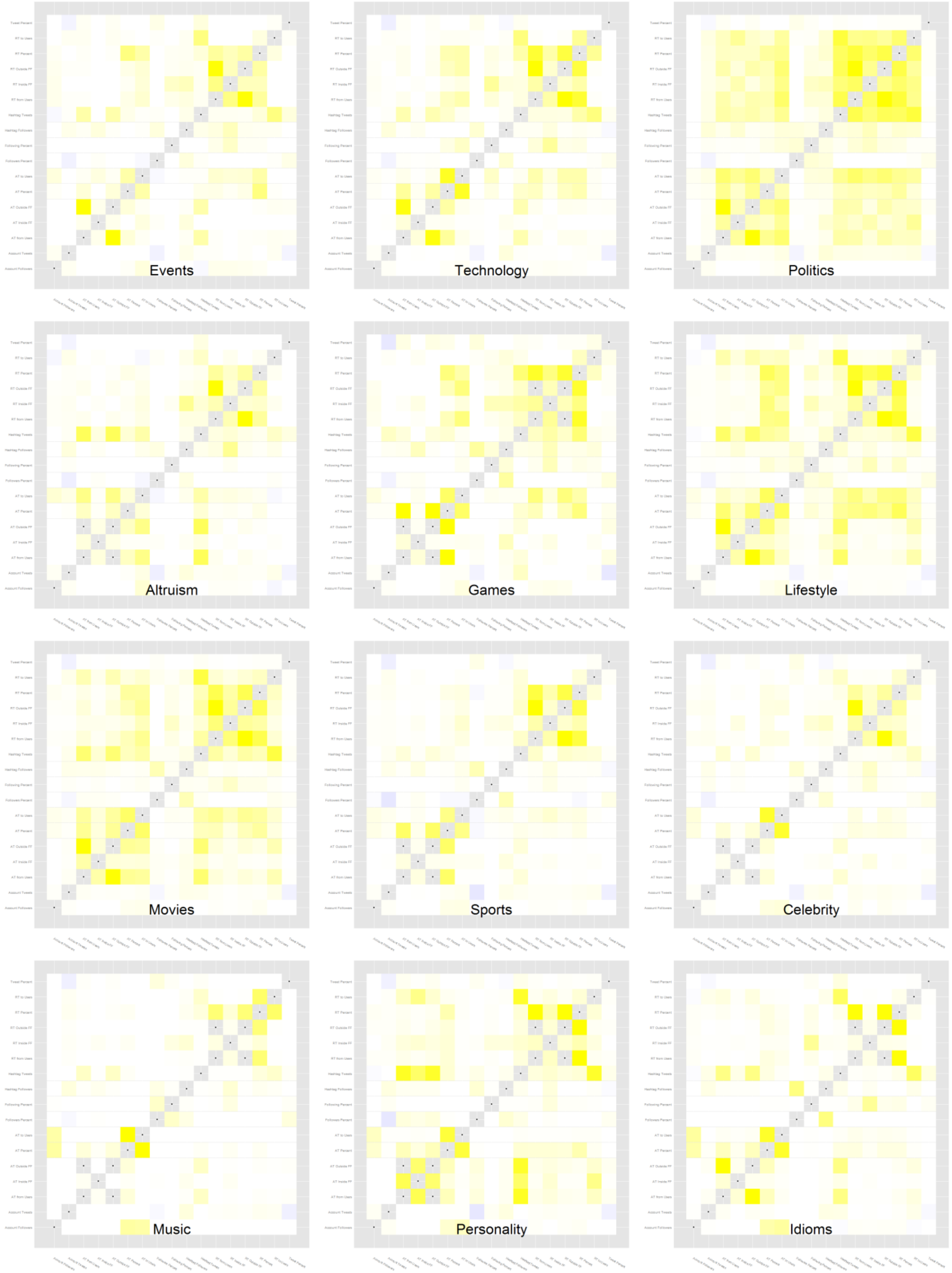

Figure 5 Correlation plot (p<0.001) for the following arrays: tweet percent; RT to users; RT percent; RT outside FF; RT inside FF; RT from users; hashtag tweets; hashtag followers; following percent; followers percent; AT to users; AT percent; AT outside FF; AT inside FF; AT from users; account tweets; account followers.

8. Conclusions

RT and AT-message networks present different levels of user interconnectivity and suggest the participation of different publics. Retweeted news that raises public awareness (i.e. Altruism) or is motivated by personal interest (i.e. Technology) presents a level of user interconnectivity which is considerably higher than average (5% and 6%, respectively). Idioms and Personality hashtags spread through the Twittersphere without resorting to the networks of individual users, as only 3% and 5% of the retweets within these topics involved interconnected users. We also found that AT-messages that included a hashtag present no significant variation in regard to the interconnection of users as followers and followees.

These figures indicate the importance of factors other than the network of followers and followees to the spread of messages. In order to assess the importance of the Twitter Trending Topic section, we collected two-thirds of our data after the hashtag appeared in Twitter Trending Topics, while one-third of dataset was archived without ever having made it to Twitter Trending Topics. We hypothesized that hashtags which featured in Twitter Trending Topics would have a lower rate of user interconnectivity, because the retweets were broadcasted in the Trending Topics section. However, we found no significant deviation between hashtags classified in the same topic which featured in Twitter Trending Topics and those that did not in regard to the number of retweets between interconnected users.

The statistically lower AT-message connectivity in Games, Celebrity and Sports indicates that users commenting on these topics are more likely to message other users that are not part of their personal network. We understand these figures to reflect game users' habit of setting up the game platform to post their results and scores on their Twitter account. Given that user scores can be updated very frequently, the hashtag data might include a disproportionate number of game statistics instead of a proper user to user conversation. As a result, Games hashtags and keywords have less conversational features.

Celebrity hashtags often contain infotainment news or users' comments on celebrities private lives. The celebrity addressed by Twitter users rarely answers the messages, thus shaping a network in which many users mention a specific user who answers no one and fosters no conversation. Sports hashtags fall in between Celebrity and Games. It often includes the latest developments on sports competitions and is not intended to start a conversation. Although Celebrity AT-messages mirror the activity of fan clubs centered around celebrities, Sports AT-messages suggest a group of users who are not united, but divided, by teams. Therefore, Sports-related tweets can feature sports celebrities in a way similar to Celebrity messages, while also including up-to-the minute headlines on sports scores.

Romero et al. (2011) defined Twitter-Idioms as tags that did not include any name of a person or a location. Even though this definition is broadly consistent with the one used throughout this paper, our method of classification based on RT-connectivity found hashtags that included the name of a person and/or a location, e.g. brazilwaits4bustinjieber, and which are consistent not with the results found in Music or Celebrity, but instead with the Twitter-Idioms group. Another difference is that the Twitter-Idioms definition of Romero et al. (2011) contained the names of days in a concatenation similar to Idioms-hashtags, including Twitter-invented holidays like MusicMonday or FollowFriday. Nonetheless, the results of our classification based on RT-connectivity show that these hashtags are shared by users with very different levels of interconnectivity, and therefore an alternative group was created to gather these hashtags and keywords (Events).

Lastly, the classification of Twitter hashtags based on RT-connectivity is consistent with the principles of hard and soft-news (Tables 2 and 5). These results point toward the possibility of automatized content classification of Twitter messages based on the interconnectivity of the users who sent and received RT and AT-messages. The results also show that retweet reliance on the network of followers and followees is relatively low, thus suggesting that Twitter users are relying and browsing other networks.

References

Bakshy, E., Hofman, J., Mason, W., & Watts, D. 2011. Everyone's an influencer: quantifying influence on twitter. In Proceedings of the ACM international conference on Web search and data mining, Hong Kong.

Bourdieu, P. 1998. *On television and journalism*. London: Pluto Press.

Franklin, B. 1997. *Newszak and News Media*. London: Arnold.

Galtung, J., & Ruge, M. H. 1965. The Structure of Foreign News: the presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research*, 2(1), 64-90.

Groth, O. 1928. *Die zeitung; ein system der zeitungskunde Journalistik*. Mannheim: J. Bensheimer.

Harcup, T., & O'Neill, D. 2001. What is News? Galtung and Ruge Revisited. *Journalism Studies*, 22, 276-280.

Huberman, B., Romero, D., & Wu, F. 2009. Social networks that matter: Twitter under the microscope. *First Monday*, 14 1.

Johnson, T. J., & Kayer, B. K. 2004. Wag the Blog. *Journalism and Mass Communication Quarterly*, 813, 622-642.

Kwak, H., Lee, C., Park, H., & Moon, S. 2010. What is Twitter, a social network or a news media? In Proceedings of the 19th international conference on World Wide Web, New York, USA.

Newman, M. E. J., & Park, J. 2003. Why social networks are different from other types of networks. *Physical Review E*, 683.

Romero, D., Meeder, B., & Kleinberg, J. 2011. Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. In Proceedings of 20th ACM International World Wide Web Conference, Hyderabad, India.

Schudson, M. 2011. *The sociology of news*. New York: W.W. Norton.

Tunstall, J. 1971. Journalists at work. London: Constable.

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. 2011. Who Says What to Whom on Twitter? New York: Yahoo! Research.