Event Diffusion Patterns in Social Media

Minkyoung Kim, Lexing Xie, Peter Christen
Research School of Computer Science
The Australian National University
Canberra ACT 0200, Australia
{minkyoung.kim, lexing.xie, peter.christen}@anu.edu.au

Abstract

This study focuses on real-world events and their reflections on the continuous stream of online discussions. Studying event diffusion on social media is important, as this will tell us how a significant event (such as a natural disaster) spreads and evolves interacting with other events, and who has helped spreading the event. Tracking an ever-changing list of often unanticipated events is difficult, and most prior work has focused on specific event derivatives such as quotes or user-generated tags. In this paper, we propose a method for identifying real-world events on social media, and present observations about event diffusion patterns across diverse media types such as news, blogs, and social networking sites. We first construct an event registry based on the Wikipedia portal of global news events, and we represent each real-world event with entities that embody the 5W1H (e.g., organization, person name, place) used in news coverage. We then label each web document with the list of identified events based on entity similarity between them. We analyze the ICWSM'11 Spinn3r dataset containing over 60 million English documents. We observe surprising connections among the 161 events it covers, and that over half (55%) of users only link to a small fraction of prolific users (1%), a notable departure from the balanced traditional bow-tie model of web content.

1 Introduction

There are massive and continuous streams of digital chatter being generated by mainstream news, blogs, social networks, and other online content. A significant portion of this chatter is driven by news-worthy, real-world events. This paper focuses on capturing such events in online media, and studying the network properties and dynamics among different events involving hundreds of thousands of online users. Many real-world events quickly spread worldwide, some immediately affect the political, economic and social lives of millions (e.g. the Blitz in London during WWII), some bear long-term cultural and ideological influence (e.g. The Declaration of Independence), and some show a significant role in both (e.g. the 1933 recession and subsequent changes in finance practices). Large collections of real-time content have only recently become available, enabling event tracking at scale. The significance of event tracking can been seen at three different levels: to quantify the extent and evolution of real-world events; to reveal the connections between different events; and to anticipate the short-term effects and long-term changes they will incur.

Existing solutions to online diffusion tracking have taken several angles. Information diffusion can be defined based on shared keywords or similar text between documents (Gruhl et al. 2004; Adar and Adamic 2005), on hyperlinks (Leskovec et al. 2007; Cha, Pérez, and Haddadi 2009), on network-specific actions such as retweets or hashtags for microblogs (Kwak et al. 2010; Romero, Meeder, and Kleinberg 2011), or on shared quotes called *meme phrases* (Leskovec, Backstrom, and Kleinberg 2009). The operational definition of an event has included hashtags manually grouped into broad topical themes (Romero, Meeder, and Kleinberg 2011), or reflections of events in meme phrases (Leskovec, Backstrom, and Kleinberg 2009). More recently, (Becker, Naaman, and Gravano 2011) designed a two-step approach to first cluster the input Twitter stream and then perform event versus non-event classification on clusters. None of these event definitions is about a particular real-world event; an instance of an event is usually known soon after its onset on the microblog stream which may or may not contain meme phrases, and it can have a complex and evolving relationship with hashtags.

We propose a novel method for finding all event-related documents across diverse media sources. This method aims at capturing online discussions about a particular real-world event, and it achieves this goal by harnessing Wikipedia for a registry of important events, representing events and documents with journalism-inspired features. This operational definition of real-world events allows us to quantify the volume, dynamics, and interactions among events. We use document hyperlinks in the main content to generate an accurate citation network. Such an event representation allows us to observe two different overlays on top of the linked documents – a network of users and another network of events (illustrated in Figure 1). We analyze the ICWSM 2011 Spinn3r dataset, with over 60 million English documents covering a one-month period in early 2011. We observe that hyperlinks across different event-related documents account for the majority of the total links, and that such cross-event links some-

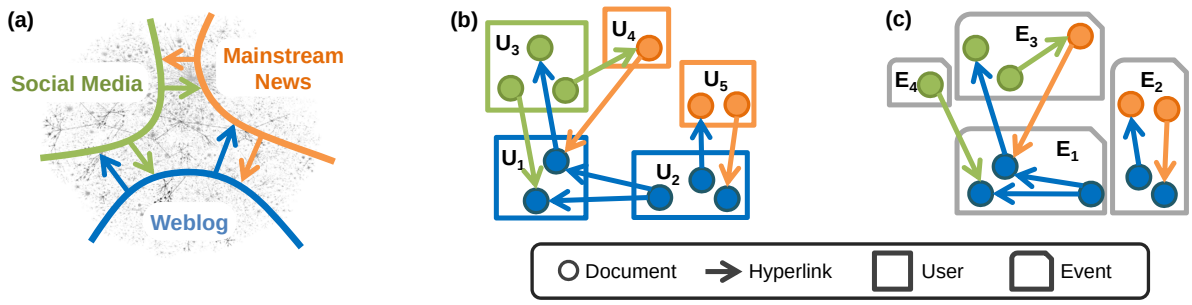


Figure 1: Event diffusion overview. (a) Three main media types that reflect real-world events. (b) The user-document network. (c) The event-document network.

times reveal surprising yet reasonable connections among events that co-evolve over time (such as the Australian Open, Queensland floods, and cricket game cancellation). We also study the user network structure with over 350K nodes. It turns out that 1% of the most productive users contribute over 40% of the content. The strongly connected core in the largest user network is much smaller than that of the well-known bow-tie model of the web. Such observations, to the best of our knowledge, is seen for the first time.

In the rest of this paper, Section 2 proposes a method for identifying real-world events, Section 3 presents our processing method and network statistics, Section 4 analyzes the diffusion patterns and interconnections between events, Section 5 observes the user networks by layering them with topological properties, Section 6 concludes this study.

2 Document Labeling with Real-world Events

A common conceptual model for describing hyperlink cascades among time-sensitive documents is to look at two-level networks among documents and users (Leskovec et al. 2007). As shown in Figure 1(b), links in the document network are aggregated by their authorship, and these user-to-user links are interpreted as cascade interactions. We propose an event-based dual representation of such a hyperlink cascade, shown in Figure 1(c). Hyperlinks among documents are aggregated by the event E each document belongs to. The resulting event-to-event links reflect the interrelationships between events as events evolve over time.

There are two challenges in extracting event-to-event networks from the underlying document citation network: to identify real-world events, and to associate each document in our collection with the events that it describes. Our approach starts by learning event models from crowd-sourced online registries. We design document features motivated by journalistic practice to make it possible to compute the similarity between a document and an event. This is notably different from modeling the dynamics of meme phrases (Leskovec, Backstrom, and Kleinberg 2009) or approximating events using text clustering (Ha-Thuc et al. 2009).

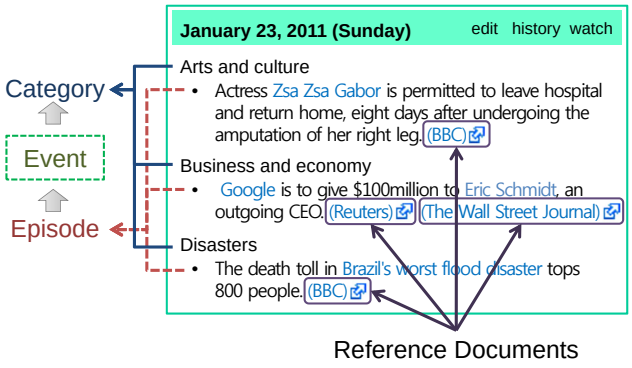


Figure 2: An example of the Wikipedia Events Portal and the associated three-level event hierarchy.

2.1 Real-world Event Identification

Choosing relevant data source of noteworthy real-world events requires some caution. Traditional news outlets such as NYTimes or the BBC are subjected to institutional and geographical biases; a social news reader such as Digg is presented by articles and not by events; news aggregation sites such as Google News have better coverage but retroactive crawling is not easy. The Wikipedia Current Events Portal[1] is a chronologically organized event registry of public interest, continuously updated and discussed by volunteers. This seems to be the best event source, despite potential selection bias of users who self-select to be editors.

Event Hierarchy from Wikipedia Event Registry We define an event hierarchy at three increasing levels of generality ($Episode \subseteq Event \subseteq Category$): $Event_i = \{Episode_j, j = 1..n\}$, $Category_i = \{Event_j, j = 1..m\}$. An *Episode* is the smallest unit of daily new happenings (e.g., mudslides near Rio), and it is possibly categorized into one general *Category* (e.g., Disaster). We argue that it is necessary to group subsequent episodes with a common subject, location, and/or real-world cause-and-effect into one *Event* as people collectively think of them as one event. For instance, some episodes are ongoing with variations (e.g., north Queensland hit by Cyclone Yasi, Brisbane River banks

[1] http://en.wikipedia.org/wiki/January_2011

break) for a short or long term, either continually or occasionally, and people think of the news as one event (e.g., 'Queensland floods').

As Figure 2 shows, an episode is mapped into each bullet point which describes a short summary of the news happening for that day. Every episode comes along with hyperlinks to one or more *Reference Documents* (circles in Figure 2) for the detailed description of the episode. We crawled all reference documents for training each episode model. The bold fonts in titles on the page are referred to as categories, and we choose five major categories that are representative across professional news agencies, and have a sufficient number of episodes over time. They are *Politics*, *Business & Economy*, *Disaster*, *Arts & Culture*, and *Sports*.

Target Real-world Events By parsing the Wikipedia event page, we built a real-world episode registry for the ICWSM'11 dataset (13th Jan to 14th Feb, 2011). We target episodes that occurred between 6th January (1 week before the beginning of our dataset period to account for the near past episode diffusion) and 30th January, 2011 (2 weeks before the end of our dataset period to observe episode diffusion for at least 2 weeks). Finally, we identified 284 episodes for this period. For the tracking of semantically meaningful events, we (the first author) manually grouped the 284 episodes into 161 events according to their subject, location, and proximity in time. We found that the manual approach is feasible and unambiguous for less than 200 events in this investigation. Automatic grouping and inter-user agreement can be topics for further investigation.

Episode Representation with Entities An episode, a unit of an event, is well defined by the "5W1H", i.e. *Who, What, Where, When, Why and How*, of journalistic practice. Note that among the five *W*s, at least three (who, where and when) directly correspond to entities, such as person name, organization, location, date and time indicators. Moreover, the rest of the 5W1H (what, why and how) often contain entities to make statements precise and credible. Therefore, we propose an episode representation using an entity vector (Figure 3(a)) whose elements consist of the TFIDF (term frequency-inverse document frequency) score (Manning, Raghavan, and Schutze 2008) of each entity extracted from the episode reference documents (circles in Figure 2).

Entity recognition in documents has been an active research area of natural language processing for over a decade. There are still challenges being actively tackled by the community, and efforts to date have produced high-quality tools. We conducted entity recognition by using the OpenCalais API which provides up to 116 types of entities (from *Anniversary* to *Voting Results*). We extracted 4,411 unique entities (also, using entity resolution techniques as described in the next section) for 284 episodes from the crawled reference pages, and generated both 4,411-dimensional entity vectors and their centroid for each episode.

2.2 Document Labeling with Identified Episodes

We also represent each web document as an entity vector with the same dimensions as the episode vectors. We then use the vector-space model for classifying documents into

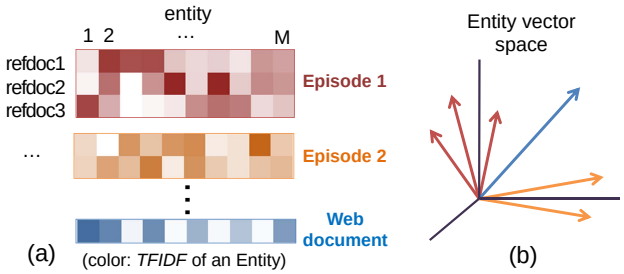


Figure 3: (a) Episode representation with entity vectors. (b) Classification of a web document with vector space model.

episodes by calculating the similarity between document vector and episode class vector (more precisely, the centroid of episode vectors) as shown in Figure 3(b); the most similar centroid vector specifies the most similar episode class. Each document can potentially be labeled with none, one, or more episode classes. For simplicity, we assign up to one episode class per document only in this work.

Entity Resolution An entity name can occur in many different ways among web documents, resulting in multiple dimensions for the same entity. For example, in our data, collection we identify nine name variations for Tunisias former president Zine El Abidine Ben Ali, including Zine Al-Abedine Ben, Zine Al-Abedine Ben Ali, and Zine Al-Abdine Ben Ali. To alleviate this problem, we employ approximate string matching techniques to cluster similar entity names. Such techniques are commonly used in entity resolution and data matching to identify similar strings that refer to the same entity.

We investigate four techniques that have been found to be effective for matching names (Christen 2006): (1) edit distance, which counts the number of character edits (insertion, deletion, and substitution) to convert one string into another; (2) Jaccard distance based on bigrams, which counts how many bigrams (sub-strings of lengths two) two strings have in common; (3) Winkler, which is a comparison technique specifically for matching surnames; and (4) the longest common sub-string approach which recursively extracts the longest sub-strings two strings have in common and then counts the number of characters in these common sub-strings. Each of these comparison functions returns a normalized similarity value between 0.0 (for two totally different strings) and 1.0 (for two strings that are the same).

Evaluation of Document Labeling For the evaluation of document labeling, we use the crawled reference documents (653) of 284 episodes as ground truth. We divided them into 538 documents for training 284 episodes and 115 documents for testing. The baseline classification accuracy based on cosine similarity between document and episode class vectors is 68%. The entity resolution brought the improvement of the accuracy up to 74% using the best-performing Winkler string matching technique.

Since not all documents in our collection are related to the identified episodes, we need to choose a similarity threshold value (τ) that filters out documents corresponding to none of the known episodes. We empirically determine the threshold

Table 1: Result of document labeling

Document Labeling	Types of documents	Document count	%
Labeled	(a) documents containing episode-related entities: (b)+(e)	2,838,692	68.60%
	(b) documents whose similarity scores of episodes are over the threshold (τ) within (a)	2,254,049	54.47%
Unlabeled	(c) documents with no content or not supported languages by entity recognition API	341,634	8.26%
	(d) documents containing no episode-related named entities	957,957	23.15%
	(e) documents whose similarity scores of episodes are under the threshold (τ)	584,643	14.13%
Total	(b)+(c)+(d)+(e)	4,138,283	100.00%

value as 0.14 which maximizes the F1 (the harmonic mean of precision and recall) score with a multi-label classification evaluation metric (Tsoumakas and Katakis 2007). As a result, we labeled the non-isolate documents (4,138,283) (Section 3.1); a subset of 2.8 million documents have sufficient text containing at least one episode entity, and within this subset, 2.2 million have an episode label with a confidence score above the threshold ($\tau = 0.14$). Details are shown in Table 1.

3 Basic Statistics of Document and User Networks

Our analysis and observations are based on the ICWSM 2011 Spinn3r dataset. This dataset consists of over 386 million blog posts, news articles, classifieds, forum posts, and social media contents collected between January 13th and February 14th, 2011. Each document includes a timestamp, author information, language, estimated spam probability, and the HTML body with hyperlinks.

Target Dataset We focus on analyzing weblog, news articles and social media, since these are most relevant to external news events, and constitute 98.37% of the data provided. We choose to only keep posts written in English to avoid the need for (noisy) translation. We filter out duplicate documents and disregard the posts with a non-zero spam probability. Nearly 6 million documents are left after the selection.

Main Content Extraction The HTML bodies of documents contain large spam links and lengthy header/footer information, which can lead to wrong interpretations of document linkage. Thus, our data processing first needs to remove such boilerplates and keep the main document content. We use the effective *boilerpipe* library (Kohlschütter, Fankhauser, and Nejdl 2010) to achieve this goal.

3.1 Document Network Construction

Link Extraction We follow out-links in the main content of a post for diffusion tracking. Hyperlinks in the document text, however, often contain shortened URLs, masking the true identity of link destination. There are over 300 URL shortening services, and this makes it infeasible to query their distinct APIs to recover the original links. Moreover, some links have been shortened more than once, further complicating the recovery. To tackle this issue, we extract the original location from the HTTP header.

After extracting out-links, we remove self-links and out-of-scope links that connect to posts outside of the dataset. We also disregard links that point to documents created after the referring document.

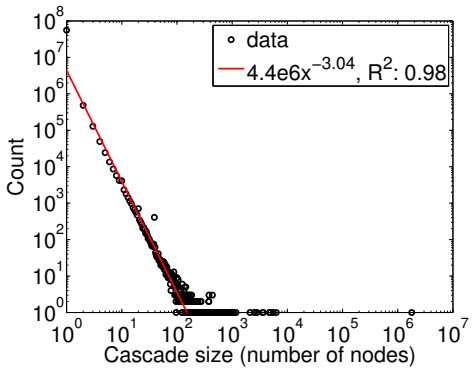


Figure 4: Cascade distribution of non-isolated documents

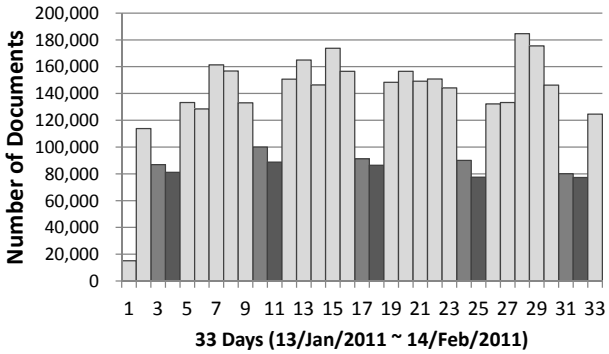


Figure 5: Daily number of documents over 1 month period. (Weekends are denoted as darker bars.)

Non-isolate Documents We filter out isolated documents which have no links to, and are not linked from other documents in the collection. The non-isolated portion has 4,138,283 documents, whose cascade size distribution is illustrated in Figure 4, showing a heavy tailed distribution. This accounts for 6.9% of the original 60 million documents. This is due in part to the majority of documents isolated. Note however that this is not a low percentage compared with the literature, where only 2% of 2.2 million blog posts are not isolated (Leskovec et al. 2007). In fact, our higher percentage results from links between three different types of document sources (news, blogs and SNS) based upon a wide range of content types of the ICWSM'11 dataset. Figure 5 shows the daily number of posts over the 1 month period and there is a clear seven-day periodicity: less documents are posted on weekend than on a weekday (the document size of the first day is originally small in the dataset provided).

Table 2: Identified Users on Social Media

Media Type	Site	User Count	% of Each Site
Blogs	blogspot.com	83,589	29.33%
	wordpress.com	18,490	6.49%
	tumblr.com	5,648	1.98%
	typepad.com	2,999	1.05%
	livejournal.com	7,001	2.46%
SNS	facebook.com	161,802	56.77%
	twitter.com	2,841	1.00%
	posterous.com	715	0.25%
	flickr.com	1,932	0.68%
Total		285,017	100.00%

3.2 User Network Extraction

To obtain underlying user networks as illustrated in Figure 1(b), the identity of users and their created document information are essential.

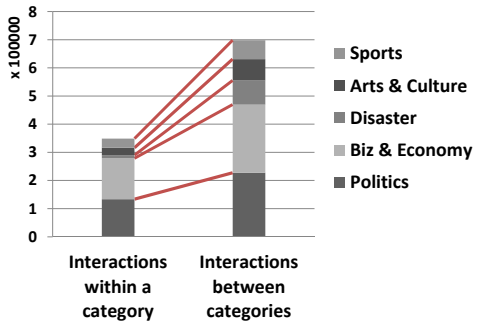
User Identity There is no universally valid user information, due to the diverse sources that the ICWSM'11 Spinn3r dataset draws from. We chose five major blog domains and four major social media domains (as shown in Table 2) as we can write regular patterns for extracting user identity from each. We also regard the second-level domain of news sites (e.g. nytimes.com, bbc.co.uk) as the unique identifier of a mainstream news agency. There are 2.2 million documents (53% of the total non-isolated documents and 76% of the largest cascade of the non-isolated documents) with known user ids. This method generates a significantly large set of users, and is consistent with prior blog user extraction methods (Cha, Pérez, and Haddadi 2009).

User Network We make a directional link from user U_1 to U_2 if there is a hyperlink from at least one document that U_1 wrote pointing to another document of U_2. The link weight is set as the number of accumulated links. This yields the underlying user networks. As a result, we identified over 350K users and their largest connected network consists of about 310K users. Details are explained in Section 5.

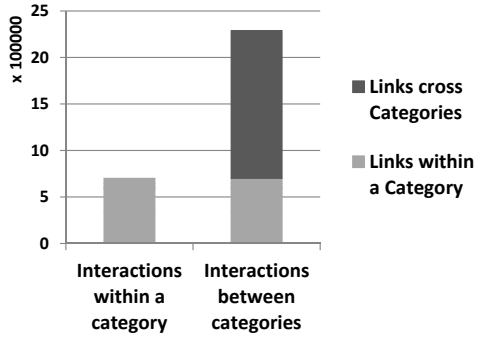
4 Event Diffusion Analysis

We extracted non-isolated documents and labeled them with the identified episodes in Section 2.2. Based on the labeled non-isolated documents, we observe the linkage patterns of episode-to-episode network and aggregate the network into an event-to-event network as is shown in Figure 1(c).

Event-related Document Network Topology We build document networks separately for each category and also generate networks of all categories combined. Figure 6(a) compares the document volumes in each category-specific cascade (different shades) versus the largest cascade across all categories (700K+ documents, bar on the right); we can see that not only do *Politics* and *Business & Economy* have more documents than the other categories within the largest cascade of all categories combined (right), the largest cascades of these two categories are also larger than that of the other categories (left). Figure 6(b) illustrates the number of links between documents which belong to the same category (left), and the different categories (right).



(a) Document volume of the largest cascade of each category accumulated (left), and of all categories combined (right): categories are shaded.



(b) Linkage volume of the largest cascade of each category accumulated (left), and of all categories combined (right).

Figure 6: Comparison of (a) document volume and (b) linkage volume within (left of (a) and (b)) and across categories (right side of (a) and (b)).

We observe that the largest cascade of all categories combined is about twice as large as the sum of the largest cascades of each category. Its linkage volume is about three times larger than the total linkage volume of the five largest cascades. This means that about 50% of documents of the all-category network are linked to documents of a different category, and there are about twice as many cross-category citations. In other words, the interactions between different categories is dominant in the largest event diffusion. Thus, we look into the linkage patterns between events across all categories in detail.

Event Linkage Patterns Figure 7 shows the normalized linkage patterns between events where the value of element (i, j) of the matrix is defined as the number of linkages from $Event_i$ to $Event_j$ divided by the total number of linkages of $Event_i$. The matrix is rearranged to group events based on the category and expressed in the gray-scale map where black is the maximum value (1.0) and white is the minimum one (0.0). Note that some dark-gray vertical lines are found in the normalized linkage matrix. These refer to some particular events that get linkages equally from most of the categories, and this causes event diffusion across categories. From our results, those events are the US banking crisis (E_{80}), Somalia pirates (E_{13}), 2011 AFC Asian Cup (E_{156}),

Table 3: Clustered events by normalized cut on the largest cascade of all categories combined

	Politics	Business & Economy	Disasters	Arts & Culture	Sports
Cluster1	0.0%	0.0%	20.0%	10.0%	70.0%
Cluster2	14.8%	34.6%	23.5%	21.0%	6.2%
Cluster3	54.5%	9.1%	18.2%	18.2%	0.0%
Cluster4	61.5%	7.7%	30.8%	0.0%	0.0%
Cluster5	53.1%	24.5%	12.2%	8.2%	2.0%

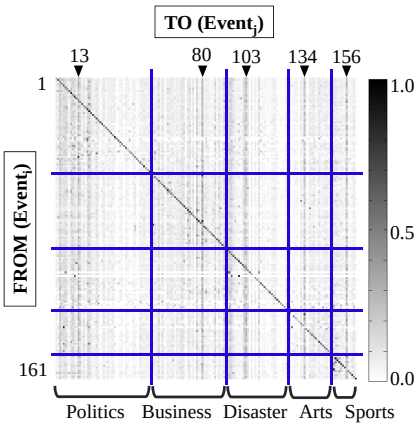


Figure 7: Normalized linkage patterns between events. (E_{13}-Somalia pirates, E_{80}-the US banking crisis, E_{103}-the Haiti earthquake anniversary, E_{134}-Movie awards, E_{156}-2011 AFC Asian Cup)

Movie awards (E_{134}), and the Haiti earthquake anniversary (E_{103}), in decreasing order of darkness.

From the observation of inter-category linkages in Figure 7, we investigate how closely events in different categories are connected by applying a clustering algorithm, the normalized cut on the network (Shi and Malik 2000). Table 3 shows the distribution of events in the clustering results by normalized cut. It is difficult to find one-to-one matching between clusters and categories, which reconfirms that events are tightly connected across categories. The events in cluster 1 are about the 2011 Australian Open, the flood in Australia, a cancellation of a cricket game, a victory of England in the 2010-2011 cricket series in Australia, and England hooligans on trial. Also, interestingly, Cluster 1 has events mostly related to Australia and England, Cluster 3 to Israel and Iran, and Cluster 4 to Pakistan and India. It shows that the contexts of events are correlated with each other across categories and clusters seem to reveal general geographical relationship among events.

Figure 8 shows a scatter plot of document in-degrees versus out-degrees for each of the five categories, and across all categories. As the figure shows there are no strong correlations between in-degree and out-degree size, which means that a document which is largely cited does not necessarily cite a lot, and vice versa.

This section discussed the interconnection patterns among events and event categories. The finding can serve as parameter estimates for the diffusion rates between events or categories, which is one of the topics for future study of event diffusion modeling.

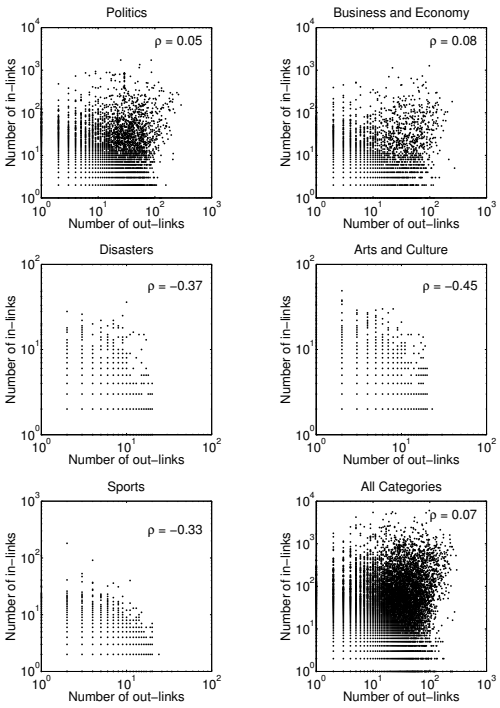
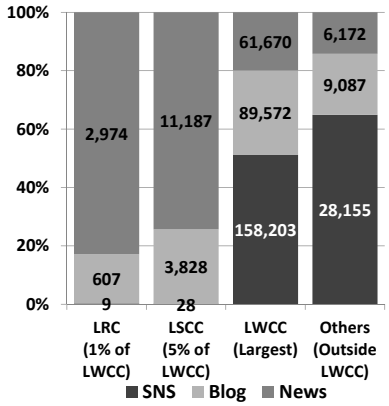


Figure 8: The scatter plot of in and out degrees of documents for each category and all combined categories

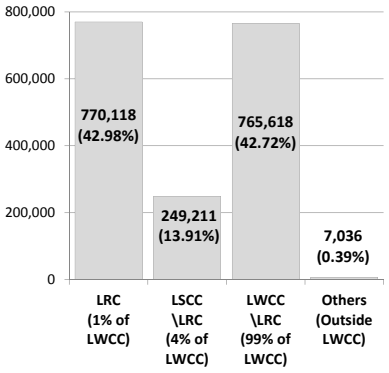
5 Underlying User Network

We analyze the user network at two different levels. The first one is a layered approach which looks into the largest connected network first, and extracts the next largest and more strongly connected network. Consequently, we extract the three major layers, namely, *LWCC* (the largest weakly connected component), *LSCC* (the largest strongly connected component), and *LRC* (the largest reciprocal core) (Wolfe 1997). Note that there is a hierarchy among the three layers: LRC is the subset of LSCC which is the subset of LWCC ($LRC \subseteq LSCC \subseteq LWCC$). The second approach is to analyze the network at a macroscopic level in order to obtain global structures and linkages within a network by breaking it into six meaningful pieces based upon the Bow-tie model (Broder et al. 2000).

Three Layers of User Network In order to discover the connections between document and user networks, we extract different levels of user networks in terms of strength of citation relationships. Figure 9 shows that the users in LRC, whose size is only 1% of LWCC (310K), create almost half (42.98%) of all documents in the largest cascade (1,791,983), while the remaining users of LWCC produce almost the same number of documents (42.72%). The users (40K) who are outside the LWCC create only 0.39% of the documents in it. Although LWCC consists of a large proportion of blogs and SNS users (about 80% in total), LRC (1% of LWCC) is mostly news media which contributes to half



(a) User size of each layer of user network and its source media type distribution



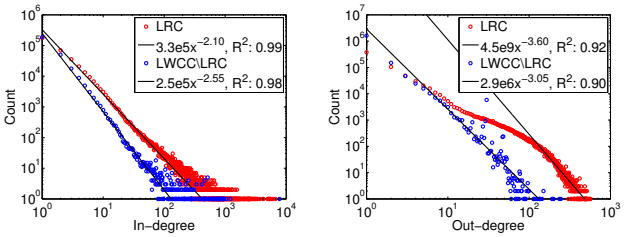
(b) Document size in the largest cascade created by each disjoint user layer

Figure 9: User layer distributions by media types (a) and document distributions created by each layer (b).

of the largest event diffusion. One possible interpretation is that with regard to public social events, individuals tend to refer to authoritative media even though there are easily accessible web posts created by micro/macro blog users.

1% vs. 99% of Users in LWCC We look into the structural patterns of document networks created by each user group. As Figure 10 shows, both the in and out degrees of the largest cascade documents created by 1% user group (LRC) of LWCC are much larger than that of the remaining 99% of user group, and there is a tendency that 1% of users receive more in-links than the other 99% users. Also, both cases tell us that documents are more likely cited than to cite.

Figure 11 shows three citation plots corresponding to each disjoint user group (top:LRC only, middle: LWCC without LSCC, bottom: LSCC without LRC). As the figure shows, LRC users make the highest number of citations to different event-related documents compared with the other groups. This fact can be thought as LRC users discover connections between different events, and these citations are observed to other users, which may contribute to the wide diffusion of events.



(a) Indegree document counts (b) Outdegree document counts

Figure 10: In and outdegree distribution of documents created by 1% versus 99% users of LWCC

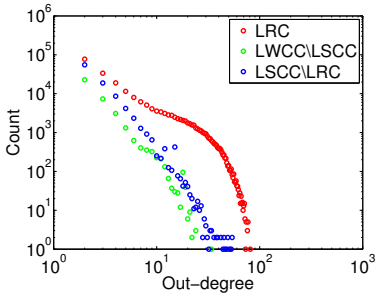
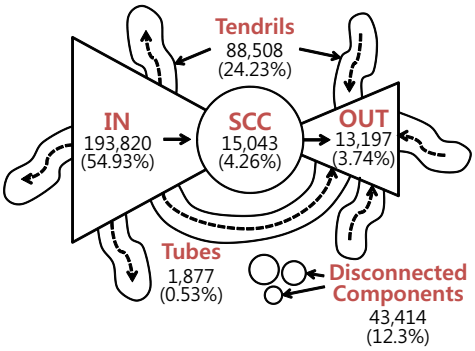


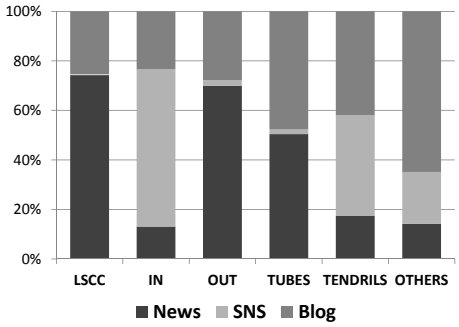
Figure 11: The number of citations to different events in a single document created by each user group (color)

Structural Properties by Bow-tie Model The user network is based upon the citation relationships, so a user citation network itself reflects document citation structure at an abstract level. For the observation of global structures of event diffusion made by user networks, we use the bow-tie model which represents the macroscopic structure of linkages. The bow-tie model breaks the whole network into six parts. The first part is a central core, *LSCC* (the largest strongly connected component). All of its nodes can reach one another along directed links at the heart of the network. Secondly, *IN* consists of nodes which can reach the LSCC, but cannot be reached from it. The third part *OUT* contains posts which are accessible from the LSCC without citations back to it. Then there are *TENDRILS* which consists of nodes that can reach either *IN* or *OUT*, and *TUBES* that link between *IN* and *OUT*, but both components have no directly links to and from LSCC. The last part consist of disconnected components (*DISC.*) outside of the largest connected component.

Figure 12(a) contains the number and percentage of users in each component. Compared to web structure (Broder et al. 2000), we can see that the size of the central core (LSCC) is small (4.3%) compared to that of the web structure (27%). Our user citation network has a much larger *IN* component than the *OUT* component (54.9 vs 3.7%), whereas both are around 21% for the web. Such a surprising small core and the imbalance of *IN* and *OUT* is partially explained by Figure 12(b). The small LSCC consists mostly of news media (74.37%) and blogs (25.45%). Note that *IN*, accounting for more than half of all users itself, is occupied by a majority of social media users (64%). The *TENDRILS* also have a significant social media presence (41%). Blog users are



(a) User Citation network as a bow-tie model



(b) Proportions of users by media types

Figure 12: (a) Macroscopic structure of the user citation network presented as a skewed bow-tie. (b) Media type distributions of each component in the bow-tie model.

positioned at all over the part with minimum 20% of proportions. The majority of SNS users tend to cite to news media, but they are not cited back from it, while blog users have bidirectional interactions with news media. Also, notable is the dominance of blog users in *DISC*, suggesting that blogs still are popular forums for things out of main-stream interest, or of non-news items. Overall, news media seem to play a central role to diffuse events over the whole web.

6 Conclusion

We present event diffusion patterns across different types of social media in terms of document network, user network, and the connections between the two networks.

First, events are interrelated with each other across all five categories (*Politics*, *Business & Economy*, *Disaster*, *Arts & Culture*, and *Sports*). This tells us that all categories need to be examined for tracking event diffusion, and also that widely spreading events have influence on wide areas of our society, namely politics, business, arts, and sports.

Second, there is a small proportion of reciprocal citation relationships between social media users which is only 1% of the largest weakly connected component (LWCC). This 1% of user group (LRC) creates half of the largest cascade of documents while 99% users of LWCC produce almost the same size of documents in the cascade. In addition, one-way relationships are more widely found between SNS users and news media than blog users and news media.

Finally, the documents created by LRC users are both cited and cite much larger documents than those generated by the remaining users of LWCC. Also, LRC users make citations to a larger number of different events in a single document than other user groups, which possibly contributes to the wide-spread across categories by showing the unexpected relationships between events to other users, and consequently, to massive diffusion of events.

Our analysis proposes approaches for real-world event tracking across different types of media. We expect that this work would shed light on an analysis of event diffusion patterns on the Web. As future works, one topic is to improve real-world event identification and model event diffusion based on this empirical study.

References

Adar, E., and Adamic, L. A. 2005. Tracking information epidemics in blogspace. In *Int'l Conf. on Web Intelligence*.

Becker, H.; Naaman, M.; and Gravano, L. 2011. Beyond trending topics: Real-world event identification on Twitter. In *ICWSM*, 438–441.

Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A.; and Wiener, J. 2000. Graph structure in the web. *Computer networks* 33(1).

Cha, M.; Pérez, J.; and Haddadi, H. 2009. Flash floods and ripples. In *3rd Int'l Conf. on Weblogs and Social Media*.

Christen, P. 2006. A comparison of personal name matching: Techniques and practical issues. In *Workshop on Mining Complex Data, held at IEEE ICDM, Hong Kong*.

Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. 2004. Information diffusion through blogspace. In *WWW*.

Ha-Thuc, V.; Mejova, Y.; Harris, C.; and Srinivasan, P. 2009. Event intensity tracking in weblog collections. In *ICWSM*.

Kohlschütter, C.; Fankhauser, P.; and Nejdl, W. 2010. Boilerplate detection using shallow text features. In *WSDM*.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is Twitter, a social network or a news media? In *WWW*.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *SIGKDD*.

Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N.; and Hurst, M. 2007. Cascading behavior in large blog graphs. In *SDM*.

Manning, C.; Raghavan, P.; and Schutze, H. 2008. *Introduction to information retrieval*, volume 1. Press Cambridge.

Romero, D. M.; Meeder, B.; and Kleinberg, J. 2011. Differences in the mechanics of information diffusion across topics. In *WWW*.

Shi, J., and Malik, J. 2000. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22(8):888–905.

Tsoumakas, G., and Katakis, I. 2007. Multi-label classification. *International Journal of Data Warehousing & Mining* 3(3):1–13.

Wolfe, A. 1997. Social network analysis: Methods and applications. *American Ethnologist* 24(1):219–220.