Tweet the Debates

Understanding Community Annotation of Uncollected Sources

David A. Shamma
aymans@acm.org

Lyndon Kennedy
lyndonk@acm.org

Elizabeth F. Churchill
churchill@acm.org

Internet Experiences
Yahoo! Research
4301 Great America Parkway, 2GA-2615
Santa Clara, CA, 95054 USA

ABSTRACT

We investigate the practice of sharing short messages (microblogging) around live media events. Our focus is on Twitter and its usage during the 2008 Presidential Debates. We find that analysis of Twitter usage patterns around this media event can yield significant insights into the semantic structure and content of the media object. Specifically, we find that the level of Twitter activity serves as a predictor of changes in topics in the media event. Further we find that conversational cues can identify the key players in the media object and that the content of the Twitter posts can somewhat reflect the topics of discussion in the media object, but are mostly evaluative, in that they express the poster's reaction to the media. The key contribution of this work is an analysis of the practice of microblogging live events and the core metrics that can leveraged to evaluate and analyze this activity. Finally, we offer suggestions on how our model of segmentation and node identification could apply towards any live, real-time arbitrary event.

Figure 1: A screen capture of Current TV's *Hack The Debate* campaign. Here we see the two candidates 23 minutes and 28 seconds into the debate. Underneath them, twitter user tomwatson's comment is displayed while sgogolak's comment, the previous twitter user's comment fades away in an animation.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*Synchronous interaction; Collaborative computing*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Informaiton Systems—*Video*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing Methods*

General Terms

Human Factors

Keywords

Twitter, Debates, TV, community, multimedia, social, centrality

1. INTRODUCTION

Recently, several web applications which allow people to converse about media content have become popular online. This ranges from photo sharing websites with easy uploading from mobile devices, like Flickr [9], to micro-blogging sites where short status messages are shared and broadcast to the world, like Twitter [21]. Being a thriving social network of photo enthusiasts, Flickr's popularity has fueled much research in the multimedia community. Twitter, on the other hand, is often examined as just a social and short messaging (140 character) service where people broadcast and reply to messages. Twitter's popularity is rising; people have begun to use Twitter to discuss live events (in particular media events), which they are attending or watching on broadcast TV. Unlike other sites where we see media stored and discussed, the media is stored externally, if at all, while the conversation ensues on Twitter. This disembodied social conversation happens as people share their awareness and comments around an event. In this article, we demonstrate how the social structure and the conversational content of these short Twitter messages, *tweets*, can provide insights into the media event's structure and semantic content of the video sources they annotate.

Twitter is used to express opinion, to contribute to a trend/meme, or to react conversationally. When centered around media, one can gather trends and topics. We find Twitter coupled with live broad-

WSM'09, October 23, 2009, Beijing, China.

cast media to be particularly empowering. First, when the tweet occurs, no media is collected for annotation. This is to say, in other sites like YouTube [25] or Flickr, the media is collected and uploaded and then annotated. With Twitter, one just annotates freely with a reference to a live event which may become a media object in the future. Second, tweets can address other twitter users or just stand out on their own. So we can see what is *conversational* and what is a *comment*. Finally, we can segment and identify topics and semantics of the video based on the network and content of the tweets centered on topic. In this article, we discuss the approach, metrics and methods for finding thematic segments and summaries of the first United States Presidential Debate in 2008.

In 2008, Current TV [7], a cable and web TV news station, recognized that the Twitter community was actively communicating about live TV. They sought to bring the comments from the general population onto their live cable TV and web broadcast of the first United States presidential debate. On September 26, 2008, Current TV ran the *Hack the Debate* [8] campaign. During the debate between Senator John McCain and Senator Barack Obama, they collected all the tweets that pertained to the debate and the election campaigns. Through an optimized editorial process, the tweets were displayed on live TV in real time. New tweets were added to the bottom of the screen while a tweening animation dissolved the previous comment (see figure 1). To easily filter out the tweets which pertained to the debate, they instructed people to annotate their tweet with the text: *#current*. In their own example:

> During the debates, chime in by including "#current" in your tweet. Example: "This discussion about universal healthcare makes me want to pop some pills! #current". [8]

The tweets were filtered for content, and not every #current tweet was displayed on live TV. Current TV pulled these tweets using a customize Twitter streaming API.

2. BACKGROUND

Around 2005, "User Generated Content" (UGC) grew popular and has fueled much research in various capacities. Much of the social UGC research centers around the collection and annotation of media. We are examining a recent fusion of media and social annotation across sources: broadcast TV and Twitter. That said, our work bears some similarity to *mashup* like applications and research. Mashups are applications built using the components from several other, typically web or social, applications. Like our work, mashups also pull information from several heterogeneous sources. We first review some research of community annotation of media as well as a few consumer applications. Then we will briefly describe Twitter and its various features which we will use in our study. And, finally, we will discuss the research findings of some studies of Twitter.

2.1 Media Annotation and Sharing

Applications like World Explorer provided landscape summaries by geo-aligning tag information gathered from Flickr [1]. Beyond mashups, others sought to find semantic distance between concepts (objects, scenes) in a visual domain as found on Flickr. [24]. The temporal dimension of video proposes some difficulty when deriving semantics. Asynchronous annotation of video has become quite common on YouTube and other sharing sites. However, a flat comment left on a 10 minute video provides little to no depth of understanding what happens in the video or why it is conversational. Much work has been done academically [4] and commercially (`http://www.hulu.com/` offers its users the ability to clip a segment for sharing on Facebook) to enhance how people can annotate and share video segments. Williams et al. [23] ran a small study illustrating a prototype that allows people to share segments of TV. Despite technological difficulties, test subjects reported feeling closer to each other and engaged in deeper sharing practices. However, such wide spread usage has yet to be exhibited, let alone extend into topical or semantic extraction. Shamma et al. demonstrated that said semantics would be tied to the pragmatics of the annotating application [18]. In a later example, Shamma et al. found when sharing video in real time, users tended to wait till after the video had ended before engaging in meaningful conversation [17].

2.2 How Twitter is Used

The Twitter service is simple. A user on Twitter can post a short (140 character) text-only message to a blog-like web page. By default this 'microblog' is world readable, however, a user can choose to restrict access to specific, approved Twitter users. A Twitter user can also *follow* (subscribe to) another user's tweets. This action creates a timeline: a concatenation of all the tweets from the users that a given user follows (including their own tweets) presented in reverse chronological order. If a user follows no other user, their timeline is just their tweets.

Aside from following, the Twitter community has created, via community convention, mechanisms for communicating in this limited medium. If a user's name (for example: barackobama) appears in tweet prefixed with an @ symbol (@barackobama), that signifies a *mention*: a reply or communication directed to that user. Twitter, responding to the community convention, makes a hyperlink back to that user's personal Twitter page. Twitter users have also begun to use explicit tags to describe their posts. Since Twitter doesn't provide this interface, users prefix a hash symbol, #, to signify the 'tag' portion of the tweet. For example, if one wants to tag a post "Can't wait to see it!" with "debate08", one would post "Can't wait to see it! #debate08". Tags are often used inline in the tweet as well, such as "My #VW is so cool!".

2.3 Twitter Studies

When it comes to Twitter, the pragmatics of its usage have only begun to be discovered. Java et al. [13] found microbloggers on twitter exhibited a higher reciprocity than the corresponding values reported by Shi et al. [19] from the Weblogging Ecosystems Workshop data [2]. One way to measure reciprocity, as seen in Shi et al.'s study, is to measure friend relationships (Mary adds John as a friend and John, in return, adds Mary). Krishnamurthy et al. [14] introduced a few categories of twitter users, also by examining the follower-to-following count of users over time.

Reciprocity can also be measured in other ways. In 2009, Gilbert and Karahalios [10], inspired by Granovetter's seminal work from 1973 [11], described several ways to measure reciprocity on Facebook to predict social tie strength. For this Twitter study, we are will examine conversational reciprocity, which occurs when ever an action is returned. For example if Mary sends John a tweet and John replys back to Mary.

3. STUDY

Twittering reactions to the debate is the act of live annotation of a broadcast media event. What can Twitter activity say about the media event? Can we use these data to find important moments in the source media or generate summaries? Twittering is a form of communication. Conversationally, can themes be extracted and realigned to the (captured) media event? Can topics or higher semantics be discovered?

We hypothesize that twitter volume over time can be used to determine segmentation points as delimited by level of interest (LOI). Further, we suggest that conversation and traffic will grow as the broadcast event comes to a close. Finally, turning to content, we hypothesize that there is a correlation between Twitter traffic and the debate's transcript from the Closed Caption stream. To address these questions, we sample and analyze the tweets around the United States presidential debates.

3.1 Data Collection

To study media usage around the first Hack The Debate campaign, we needed to gather the tweets from the debate. This started with observing the Hack The Debate campaign and following how people were tweeting. We found people using the prescribed tag #current as well as two other tags: #tweetdebate and #debate08. While the latter two tags were not part of Current TV's promotion, they did relate to the live debate and were included in this study. Twitter has rate and time limits on usage of its API. When we pulled our sample in November of 2008, each search was limited to 100 tweets max returned. To get a clean sample, we built a search/crawler script which would query for all tweets with any of the three aforementioned tags for each minute of the debate. The crawler paginates through all the tweets in the search results and serializes them to disk or a database. Search results only contain tweets from the public timeline—not private and visible to everyone. Additionally, it follows the rate limit guidelines, sleeping if it issued more queries than allowed in any given time period. We sampled 150 minutes of tweets, the first 97 minutes being the actual debate airing, the remainder was captured to examine post-debate activity.

For the study, we needed supplementary information about the debate itself. We used the first debate video from C-SPAN found on YouTube [6]. We also pulled Closed Captioning data pulled from C-SPAN. On their website, C-SPAN has an interactive timeline of the debate [5] which timestamps high level topic areas:

- Opening
- Financial Recovery
- Solving Financial Crisis
- Lessons of Iraq
- Troops in Afghanistan
- Threat from Iran
- Relations with Russia
- Terrorist Threat.

The speakers during the debate were Senator John McCain, Senator Barack Obama, and was moderated by Jim Lehrer who anchors the PBS NewsHour TV show. At the time of the debates, their Twitter accounts were @johnmccain, @barackobama, and @newshour.

We sampled from a search for the three hash tags for the 97 minute debate and the 53 minutes following the debate making a total of 2.5 hours. There were 3,238 tweets from 1,160 people. There were 1,824 tweets from 647 people during the actual debate. After the debate, we found 1,414 tweets from 738 people.

For the 2.5 hours, we found 577 @ mentions. There were 266 mentions during the debate and 311 afterwards. Within the mentions, we looked for *retweets*. A retweet is where someone repeats another user's post, normally with attribution to the original poster. For example:

> RT: @jowyang If you are watching the debate you're invited to participate in #tweetdebate Here is the 411 `http://tinyurl.com/3jdy67`

The volume of retweets was very low: 24 retweets in total in our sample, 6 during the debate and 18 afterwards.

3.2 Measuring Volume

Twitter volume shows activity on its network, and, hence, can be a proxy of interest. When examined over time, areas of high and low activity, spikes and pits, are clearly visible in the traffic volume. Twitter volume increased sharply at the end of the debate. We also found the amount of @user mentions remains fairly low but did increase towards the end of the debate along with the total volume. However, the mention volume did not spike and fall as sharply as much as overall volume. Figure 2 shows the overall volume found in our sample as well as the mentions volume.

To find actual segments, we turn to the Twitter volume over time. Past research in social chat and watch systems [17] finds that people are the most conversational after a video has completed. We saw a similar effect in the Twitter data; figure 2 shows this spike upon completion of the debate. Since a debate, like most TV programming, is segmented, could we find local areas of high volume and would those maxima relate to segment boundaries. To investigate, first we define a discrete function of time in minutes which returns the sum of tweets during that minute. This function is then smoothed using a three minute sliding window, where each point is expressed as the average of itself and its two surrounding neighbors. To automatically detect peaks in the volume of Twitter activity, we apply Newton's Method [15], a simple approach for extrema detection, which detects a point of change in the slope (roots of the first derivative) of a given function: a change from a positive to a negative slope indicates a local maximum, and a change from negative to positive indicates a minimum. This approach on smoothed twitter data is somewhat sensitive to smaller fluctuations and dips in activity at small time scales. To address this, we filter the set of detected extrema to only include the outliers who are one standard deviation away from the mean, $\mu \pm \sigma$, as measured in a 21-minute sliding window to the activity volume. See figure 3. This method returns 11 segmentation markers for the 97 minute debate. We believe this method to be highly dependent on the type of media event.

3.3 Social Graph

To understand how users are connected, we first examined the Twitter network as we collected it; that is, as an undirected collection of users and tags. We assumed users would be using the #current tag and maybe one of the other two tags (#debate08 or #tweetdebate), however, when we examine tags as boundary objects between people [20], we see distinct groups with some overlap, see figure 4. The betweenness centrality [22] (vertices which occur on the many of the shortest paths between other vertices) of #current is 1.0; #debate08 and #tweetdebate scored 0.892 and 0.499 respectively. During the debate, 48.8% of the users did not use the #current tag. This shows the practice of live tweeting of this event was not limited to the Current TV Campaign.

As previously discussed, past research examined reciprocity via the following/follower relationship amongst users [13, 14]. However, as previously illustrated within the multimedia community [18, 4], conversational structures offer descriptive and meaningful media annotations. Within Twitter, the following:follower ratio is an insufficient proxy for conversation. The @user mentions, however, do *call out* attention another individual user and can elicit a response.

In our sample, the network of 577 @user mentions, see figure 5, is a directed graph of such elicit call outs. Out of all the mentions in the 2.5 hour sample, 10.23% were reciprocated. To find important nodes (people) within the network, we examined the *eigenvector centrality* (EVC) which is defined as the principle eigenvector of the adjacency matrix [3]. In short, a user will have a high EVC if they are connected to a set of users who, in turn, are con-

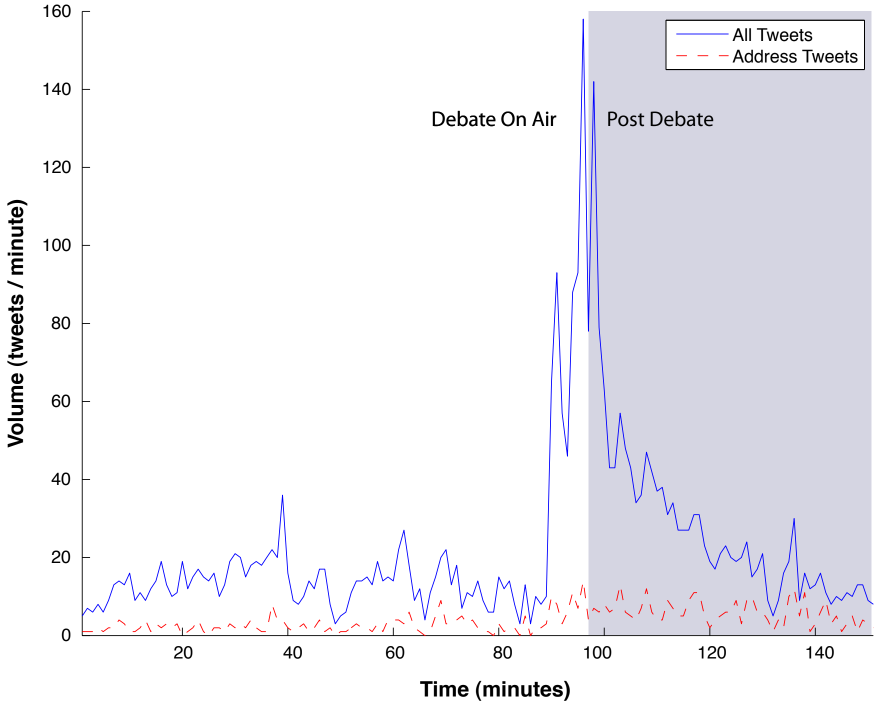


Figure 2: The volume of tweets over time sampled from the first debate time and tagged #current, #debate08, or #tweetdebate. The debate aired from minute 0 to minute 97. The swell of conversation, the shaded region, occurred mostly after the debate had ended. The blue/solid line shows the total tweets. The red/dashed line shows the volume of @user mentions.

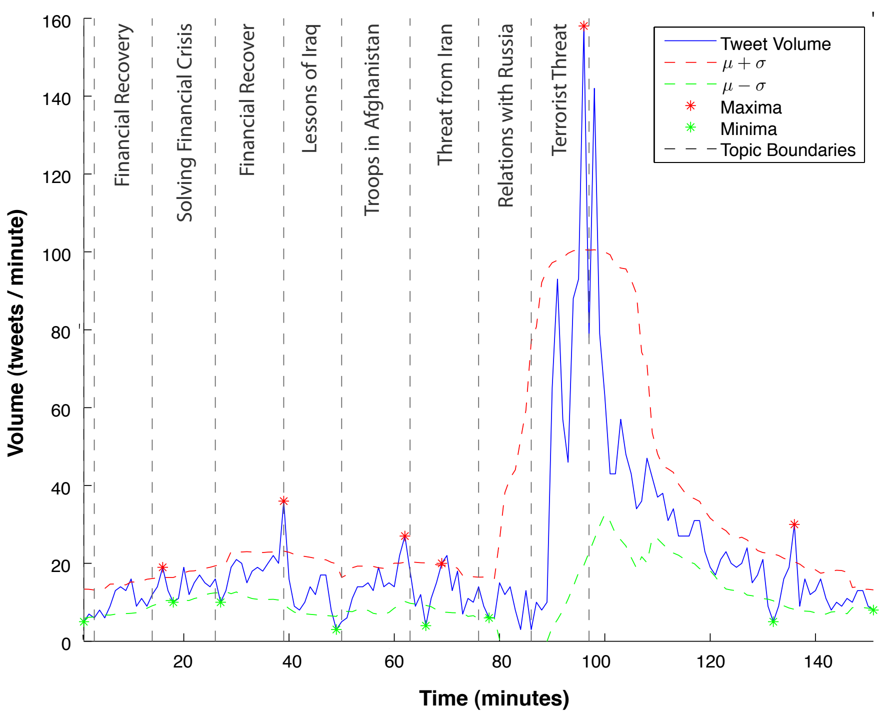


Figure 3: The volume of tweets over time sampled from the first debate time and tagged #current, #debate08, or #tweetdebate. The each vertical region represent segments as described from C-SPAN. The dotted line above and below the curve is $\mu \pm \sigma$ from a 21-minute moving window.

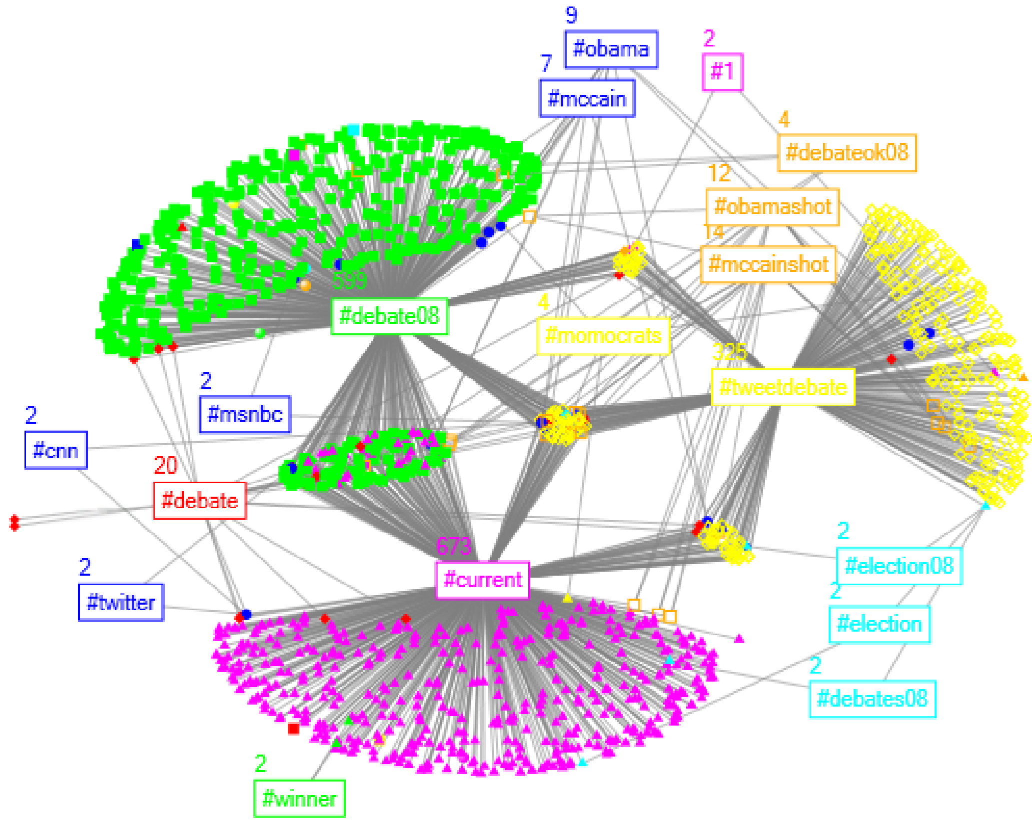


Figure 4: The network graph of all users and their tag relations from the Debate 2008 sample as seen when clustered by tags. Half of the users used the #current tag when discussing the debates during air time. The number next to each #tag denotes the degree of that node. Only tag node degrees ≥ 2 are shown for clarity.

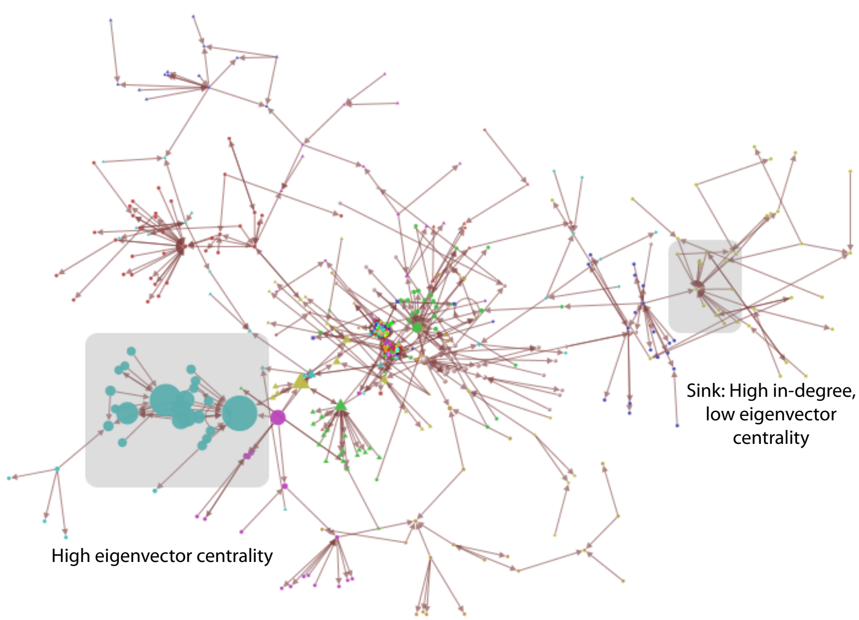
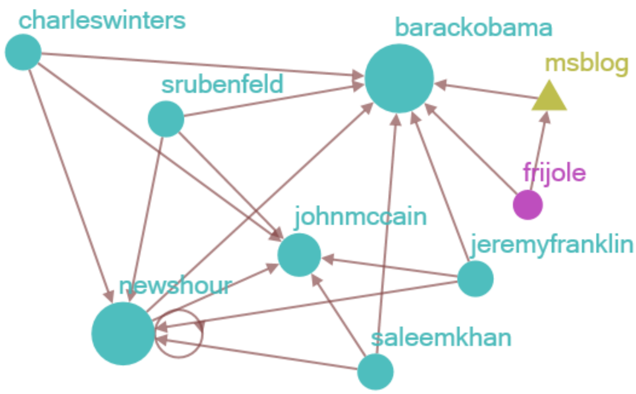


Figure 5: The directed network of Twitter Users @mentions. Larger sizes denotes a higher eigenvector centrality, shapes denote clusters. Left: a clustered region of high eigenvector centrality includes Twitter accounts from Barack Obama, John McCain and Jim Lehrer. Right: a *sink* is a node with a high in degree but low eigenvector centrality.



Twitter User	Eigenvector Centrality	In Degree	Out Degree
@barackobama	0.472	15	0
@newshour	0.427	11	5
@johnmccain	0.277	6	0
@charleswinters	0.223	0	3
@jeremyfranklin	0.223	0	3
@saleemkhan	0.223	0	3
@srubenfeld	0.223	0	3
@msblog	0.221	5	6
@frijole	0.175	0	7

Figure 6: A subgraph of Twitter Users from Figure 5 with the highest eigenvector centrality. The top three users were the three people in the debate: two candidates and the moderator. @newshour contained a self-referential tweet where they mentioned themselves.

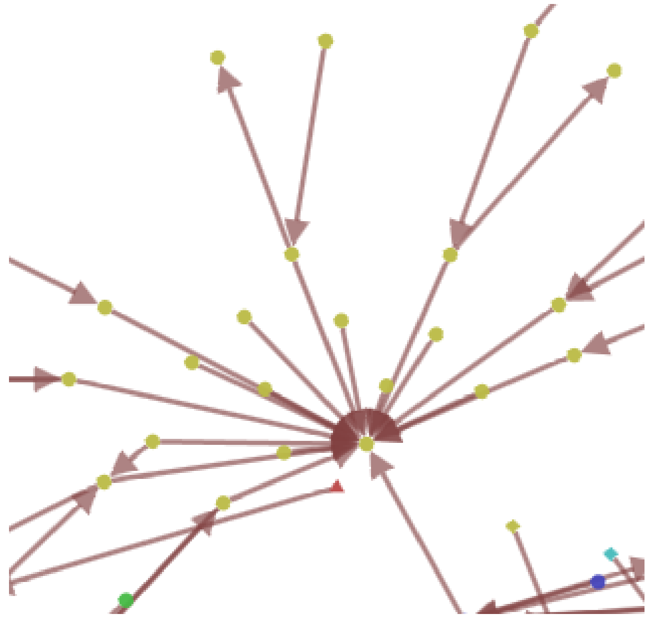


Figure 7: A cutout of Twitter users from figure 5 centered around a *sink*, a node having a high in degree and little to no out degree and low eigenvector centrality. The two predominant sinks in this network are @current, who ran the *Hack the Debate* program, and @jowyang, an employee of Forrester Research who uses Twitter as a personal, not corporately related microblog.

nected to many users. We computed the EVC using the acclerated power method [12]. Within the graph, the top three nodes with a highest EVC were the three people in the debate: @barackobama 0.472, @newshour 0.427 (Moderator Jim Lehrer), and @johnmccain 0.277. See figure 6.

Nodes with a high eigenvector centrality also possessed a high in degree, surpassed only by @jowyang, from Forrester Research who tweets personal opinions, and @current. We refer to these two nodes as *sinks* given their overall lack of out degrees and their general disconnect from the overall social graph. This is reflected in EVC of @jowyang and @current: 0.001 and 0.017 respectively, see figure 7.

3.4 Twitter Terms and Debate Dialog

Turning to content, we hypothesized that frequent terms from twitter traffic would reflect the topics being discussed. For this, we used C-SPAN's topic segments to break the debate into nine pieces. Each segment has a general category which corresponded to Jim Lehrer's questions to the candidates. For each segment, we apply a variant of the standard vector space model [16]: we count the frequency of the term in the given segment (term frequency) and divide that value by the total number of segments in which the term appears (inverse document frequency). We apply this approach separately to the closed caption transcript of the debate and the tweets posted during the segment.

Upon inspection, there appears to be some topical alignment between the tweets and the debate's closed-captioning. Figure 8 shows the top-ranked terms from each segment both from the actual debate dialog and from the captured tweets. To numerically evaluate the presence (or absence) of topic-level alignment between the debate content and the discussion on Twitter, we would like to calculate the similarity between the term vectors produced by each source; however, this introduces two problems. First, the vocabulary on Twitter varied from the rhetoric of the debate such that there was no strong alignment from inspection. More so, the divergent vocabularies introduced questions around how to handle the inverse document frequencies: different terms were used with very different frequencies across the two sources.

Second, the usage of twitter was not one of summarizing or even discussion about the debate on hand. While we found some evidence of debate topics being discussed, the communication was more so reactionary and evaluative. For example, the most frequent term during the opening segment from Twitter was "drinking"—we assume people were inventing drinking games to play along with the debate. Later in the debates we see terms like "-5" or "+2" becoming salient where people were keeping score on which candidate won which point.

4. DISCUSSION

From our hypothesis, we discovered the structure of Twitter traffic can provide insights into segmentation and entity detection, however, the correlation between content leaves further questions to be investigated. Our approach began without examining content, only looking at Twitter volume. Our method of finding segmentation points produced 11 cuts in the 97 minute debate. When we compare our method's accuracy to C-SPAN's editorial segmentation markers, we find 8 of the 11 markers were ±1 minute of the human made cut, leaving 3 false positives. If we apply a simple heuris-

	drinking	kennedy	cut	screen	senate	pulling	lousy	***georgia***	security
	candidates	plan	***earmarks***	bailout	difference	***pakistan***	democracies	video	experience
	wins	hope	compared	energy	***winning***	understand	lol	blog	9/11
	minute	hand	***tax***	economic	strategy	drinking	government	league	***safe***
	lehrer	comment	joke	power	iraq	republican	story	***russia***	country
Twitter	letting	moderator	bear	condoulo	festooned	strategy	wars	ha	tactics
	tv	tie	dollars	nuclear	fact	sounds	league	lot	attack
	mccains	home	problems	giving	war	coming	times	condescending	video
	tweet	audience	personal	freeze	giving	telling	***iran***	issue	pakistan
	moderator	republicans	american	looked	hand	story	-5	oil	bringing
Topic	**Opening**	**Financial Recovery**	**Solving Financial Crisis**	**Financial Recovery**	**Lessons of Iraq**	**Troops in Afghanistan**	**Threat from Iran**	**Relations with Russia**	**Terrorist Threat**
	presidential	street	$18	programs	funding	border	henry	***georgia***	restore
	debates	greed	requests	medicare	***winning***	taliban	kissinger	international	knowledge
	minutes	main	***earmarks***	cost	leave	prepared	contacts	putin	missile
	eisenhower	how's	gateway	eliminate	timetable	supported	preparation	ukraine	***safer***
	financial	house	loopholes	$700	violence	muddle	legitimize	***russia***	veterans
Debate	direct	package	$5,000	hard	lessons	u.s	table	world's	terms
	policy	accountable	pork-barrel	decisions	baghdad	bombing	***iranians***	offshore	focused
	news	wall	employer	agency	surge	***pakistan***	sanctions	aggression	earth
	mississippi	rewarded	business	rescue	succeed	qaeda	ahmadinejad	resurgent	billions
	university	crisis	***tax***	tom	started	army	precondition	nato	challenges

Figure 8: Top terms in rank order from the twitter stream (top) and the debate closed captioning (bottom). The segment topics were specified by C-SPAN's interactive web interface. Stems common across segments are *emphasized*. While there was little direct vocabulary overlap between stems, topically we see a match of context.

tic that disallows any cut to be made in the minute after a cut has been established, this would remove 2 of the three false positives and resulting in an accuracy of 90.9% ($\pm$1 minute) for our method. We believe this approach can be generalized across event genre's by modifying the smoothing window of 21 minutes to fit the length and pace of the event being twittered.

Beginning to examine content, we filtered the social graph to show only @user mentions in a directed network. With the mentions graph, we found the eigenvector centrality to bring the three people physically in the debate. Of course, one must be present and active in the network, which speaks towards journalism at large: politicians, TV shows, and celebrities need a Twitter presence. Interestingly, other journalists and even Current TV's twitter user had low eigenvector centrality. We wish to investigate what Twitter usage at what volume might increase a node's centrality. More so, with such a low reciprocity rate amongst direct mentions, the mention "conversations" are actually directed at people who might be unlikely to respond (in example: @barackobama and @johnmccain who are, debating and not twittering at the time). Recently, we have begun to see journalistic efforts to respond on-air to incoming tweets as well as many online live news programs display related tweets in real-time with the video.

Finally, the deep exploration of content brought more questions. The vocabulary between the two corpora (the Tweets and the debate's dialog) does introduce a standard problem in information retrieval. However, the practice of Twittering debate-like events actually tracks take away reactions and not topical discussions. We believe a vocabulary could be built across events to address both of these issues. We chose to look for extrema in the volume as a method of onset detection as our assumption was conversational volume hits a maximum towards the end of a segment. We did expect to find artifacts of topical bleeding across segments due to latency (people have to watch and twitter, which takes time). Such artifacts were not evident in our dataset. In short, our findings suggest that, over time, one could build a data set with clear mappings (both in onset detection and vocabulary) to an event genre.

5. FUTURE WORK

We have begun to explore twittering reactions to live broadcast media in the hopes to understand both the source media itself (including its structure, actors, and format) as well as the media's social conversational content. Consistent with previous research [17], we saw a large spike in twitter volume (both tweets and mentions) after the event had concluded. Upon examination of the twitter activity from the other two presidential and the one vice presidential debate, we found similar spikes after the event.

We did not find as many threaded conversation via @user mentions as we predicted, however this might be an artifact of crawling a rate-limited search API. For future studies, we plan on examining other genres from data collected via Twitter's streaming API (released June 2009). Along with using streaming data, we wish to investigate if our model of segmentation and key node identification can be used live and in real time towards any arbitrary event.

This suggests a model that can fine tuned based on real time genre identification. For real-time tracking of a live media event, the size of the social graph and the number of tweets would grow over the life time of the event and perhaps beyond. This temporal growth is highly trackable and would speak to how we measure centrality and how we find event onsets. In particular, eigenvector centrality works well for a complete social graph, but further investigation must be done to see if, in fact, it would work in real time. This speaks to the practice of retweeting. If some event types are well influenced by retweeting, closeness or between centrality might become a stronger indicator of focus at the start of an event. Then, if the social dynamic were to shift from retweeting to mentions, eigenvector centrality might prove to be a more salient metric.

Finally, we believe our approach can expanded to provide insights into non-televised events. In particular, larger and longer scale media events have no single media object to reify a collection of social Twitter data against. We wish to investigate how these larger scale events, for example tracking the whole election for six months, can be understood by examining the overall social conversation against a larger collection of associated media objects such as videos, news articles, and photos.

6. ACKNOWLEDGMENTS

We would like to thank the plurality of individuals who chatted along side us—from everyone who played *Hack the Debate* to anyone who happens to shout at the actors during a movie. Ben Clemens and Chloe Sladden were pretty cool, not that they shouted in any theaters. We also thank M. Cameron Jones, who surrendered his laptop so we could run NodeXL for half a day, along with Zach Toups and Malcolm Slaney, who provided many valuable comments. Finally, this work is made possible by the guidance and support of the Internet Experience Group at Y! Research.

7. REFERENCES

[1] S. Ahern, M. Naaman, R. Nair, and J. H.-I. Yang. World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 1–10, New York, NY, USA, 2007. ACM.

[2] The 3rd annual workshop on weblogging ecosystem: Aggregation, analysis, and dynamics. 15th world wide web conference, May 2006.

[3] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120, 1972.

[4] P. Cesar, D. C. Bulterman, D. Geerts, J. Jansen, H. Knoche, and W. Seager. Enhancing social sharing of videos: fragment, annotate, enrich, and share. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, pages 11–20, New York, NY, USA, 2008. ACM.

[5] CSPAN. Cspan debate hub. `http://debatehub.c-span.org/index.php/debate-1/`, 2008. Accessed July 2, 2009.

[6] CSPAN. First 2008 presidential debate (full video). `http://www.youtube.com/watch?v=F-nNIEduEOw`, September 2008. Accessed July 2, 2009.

[7] Current TV. `http://www.current.com`.

[8] Current TV: Hack the Debate. `http://current.com/topics/88834922_hack-the-debate/`.

[9] Flickr. `http://www.flickr.com`.

[10] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, pages 211–220, New York, NY, USA, 2009. ACM.

[11] M. Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360, 1973.

[12] H. Hotelling. Simplified calculation of principal components. *Psychometrika*, 1(1):27–35, 1936.

[13] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65, New York, NY, USA, 2007. ACM.

[14] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 19–24, New York, NY, USA, 2008. ACM.

[15] Newton's method. `http://en.wikipedia.org/wiki/Newton's_method`. Accessed July 2, 2009.

[16] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communcations of the ACM*, 18:613–620, 1975.

[17] D. A. Shamma and Y. Liu. *Social Interactive Television: Immersive Shared Experiences and Perspectives*, chapter Zync with Me: Synchronized Sharing of Video through Instant Messaging, pages 273–288. Information Science Publishing, Hershey, PA, USA, 2009.

[18] D. A. Shamma, R. Shaw, P. L. Shafton, and Y. Liu. Watch what i watch: using community activity to understand content. In *MIR '07: Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 275–284, New York, NY, USA, 2007. ACM.

[19] X. Shi, B. Tseng, and L. Adamic. Looking at the blogosphere topology through different lenses. *Ann Arbor*, 1001:48109, May 2007.

[20] C. Spinuzzi. *Tracing genres through organizations: A sociocultural approach to information design*. Mit Press, 2003.

[21] Twitter. `http://www.twitter.com`.

[22] S. Wasserman and K. Faust. *Social Network Analysis: methods and applications*. Cambridge University Press, 1994.

[23] D. Williams, M. F. Ursu, P. Cesar, K. Bergström, I. Kegel, and J. Meenowa. An emergent role for tv in social communication. In *EuroITV '09: Proceedings of the seventh european conference on European interactive television conference*, pages 19–28, New York, NY, USA, 2009. ACM.

[24] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li. Flickr distance. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, pages 31–40, New York, NY, USA, 2008. ACM.

[25] You Tube. `http://www.youtube.com`.