Twitter Usage in Indonesia

Kathleen M. Carley[1], Momin Malik, Mike Kowalchuk
Jürgen Pfeffer, Peter Landwehr

December 21, 2015
CMU-ISR-15-109

Institute for Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Center for the Computational Analysis of Social and Organizational Systems
CASOS technical report.

The author's would like to thank Brendan O'Connor for access to the decahose and support in extracting the relevant tweets. This work was supported in part by the National Science Foundation 00361150115291 with support for Twitter analysis of Indonesia and the Office of Naval Research N00014140737 with support for CMU social media analytics. Additional support was provided by the center for Computational Analysis of Social and Organizational Systems (CASOS) at CMU. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, the Office of Naval Research or the U.S. Government.

[1] Corresponding Author: Prof. Kathleen M. Carley (kathleen.carley@cs.cmu.edu), Institute for Software Research, Carnegie Mellon University, Pittsburgh, PA 15213 USA.

Keywords: Indonesia, Twitter, Padang, multi-component network, retweets, social media, spatio-temporal patterns, disaster management.

Abstract

Social media is playing a growing role in disaster management and response. Expectations are that such media can be critical for sending alerts, identifying critical needs, and focusing response. However, for social media to be used in that way, it will be necessary to understand how social media is used during normal, i.e., non-disaster periods. Herein we examine the use of a particular social media, Twitter, and assess its value for disaster management with a focus on planning and early warnings. Our focus is Indonesia, and the potential use of Twitter to support tsunami warning and response. We assess alternative collection strategies and analyze Twitter usage under normal conditions and then use this information to identify the strengths and weaknesses of this data in supporting disaster management in terms of coverage, spatio-temporal patterns, and identification of opinion leaders. We find that while one can potentially leverage Twitter for disaster management, careful collection, assessment, and coordination with official disaster Twitter sites and local on-scene Twitter opinion leaders will be critical. Guidelines for harnessing Twitter data for disaster management are provided.

Table of Contents

1 – Introduction

Government officials, first responders and the general public are increasingly looking to social media as a critical communication and monitoring tool for disaster management. Due to widespread press coverage and ease of data collection, Twitter in particular is viewed as a critical communication media for disaster management. Key advantages often touted are crowd-sourcing, speed, and the ability to access from mobile devices. Key potential disadvantages often touted include bias in user base; inaccurate, false, and out-of-date information; and reduction in access as cell-towers and electricity are out due to the disaster and may remain out for a significant period. The general discussion of the strengths and weakness of Twitter for disaster management is typically based on anecdotes and isolated case studies of usage during a disaster or in its immediate aftermath. In contrast, we are concerned with the baseline; i.e., how is Twitter normally used and what does this normal usage imply about systematically collecting and using this data for disaster management? Herein, we ask, for Indonesia in particular, how is Twitter used under normal, i.e., non-disaster conditions and what are the biases in this data particularly as they relate to coverage, spatio-temporal patterns, and the identification of local opinion leaders? How do the biases in the normal use of Twitter in Indonesia impact its utility for disaster management? Are there special considerations that must be addressed when collecting and utilizing Twitter data that increase its value for disaster management?

We do not address how Twitter is used during a disaster for disaster related communication; rather, we examine the biases in the normal use of Twitter and the way it is collected that effect its utility as a source of information for disaster management. While our data provides some guidance for coordination during disasters, we are mostly concerned with the period before the actual disaster strikes. As such, this paper is more informative about what data to collect to make sure that coordination activity is well informed than about how to coordinate. It is more about the strengths and limitations of Twitter in the Indonesian context for use in disaster management, than about how Indonesian use Twitter during a disaster. Our intent is to provide new insight into how to collect and use social media, and in particular Twitter, for disaster management with an emphasis on planning and early warning. Thus we do not focus on tweets during a disaster or its immediate aftermath, but on tweets during non-disaster periods and the biases in the data. Without understanding the biases, imposed by the data collection strategy, the technology, or the culture that is producing the tweets, one cannot use the data sensibly for planning or warning.

Why should disaster management studies be concerned with social media in general and Twitter in particular? Social media is used at all phases of disaster management, early warning, communication and organization during disasters and in their aftermath (Hossman et al., 2011), identification of critical needs in the early response (Muralidharan et al., 2011), Social media in general, and Twitter specifically, are often hailed as "the next generation" in crisis response tools (Tepstra et al., 2012; Palen et al. 2008, Palen et al 2009, Sutton et al 2008). Twitter has risen to particular prominence because its API has made searching for and storing disaster related content relatively trivial (Vieweg et al, 2010; Palen, 2008). Attention to Twitter data during a disaster increases situational awareness and supports resilience (Tobias, 2011; Palen et al. 2010). However, social media data can be difficult to use in in a disaster context. For example, in Twitter, the tweets may be inaccurate (Thomson et al, 2012), false (Tinker and Vaughan, 2010), outdated (Acar & Muraki, 2011), filled with irrelevant information (Hughes & Palen, 2009; Acar & Muraki, 2011), or being used to spread harmful rumors (Castillo et al, 201) thus increasing distrust and chaos. In general, tweets from government and from official responders are more retweeted and so such "official information tweets" can better meet the public's need to know than those by random individuals (Thomson et al, 2012).

Why is Indonesia a good venue for the assessment of Twitter for disaster management? Indonesians are strong social media consumers. Indeed Indonesia is often one of the top five nations invested in social media in general, and Twitter in particular, as will be discussed more. Indonesia is also impacted frequently, and often devastatingly, by natural disasters – tsunamis, volcanic eruptions, earthquakes, and floods (Diley, 2005; Peduzzi et al., 2009). Indonesia is also undergoing rapid socio-cultural transformations with a growing political and scientific community concerned with social media and disaster management (Philpott, 2000; Couldray & Curran, 2003), even as it plays out against a backdrop of social-media usage by terrorists and human-traffickers in the region (Eickelman & Anderson, 2003). The level of Twitter penetration, combined with Indonesia's high vulnerability to multiple natural disasters (Center for Excellence in Disaster Management & Humanitarian Assistance, 2011), suggests considering Indonesia through the lens of recent work on the use of social media as a tool in crisis response and disaster management (Carley, 2014; Landwehr & Carley, 2014). The use of social media, specifically Twitter, for disaster management has not been systematically considered for Indonesia.

In this paper, we assess the issues in using Twitter to support disaster management, particularly planning and early warning, with special attention to its use in Indonesia for tsunamis. There are three parts to this assessment. First we examine the socio-technical context to identify the trends as well as the socio-cultural-technical affordances and constraints in using Twitter. To this end we review the nature of Twitter, its use in Indonesia, the socio-cultural context and the relevant disaster context. This review is set against the context of Twitter use in disasters more generally, and the special problems of using Twitter for tsunamis. Second, we provide a description of alternative collection scenarios and assess the strengths and weaknesses of these for using Twitter for disaster management. Third, we analyze collected data to characterize the way in which Twitter is used normally in Indonesia more broadly and in Padang Indonesia in particular. This analysis is oriented around identifying the inherent biases in that data that are of particular import for disaster management specifically as they relate to coverage, spatio-temporal patterns, and identification of local opinion leaders. This overall assessment suggests a number of guidelines for harnessing Twitter data to support disaster management.

2 – Socio-Cultural-Technical Context

Based on the theory that computers should be part of the disaster management team (Carver and Turoff, 2007), we argue that social media in general, and Twitter in particular, should be part of the team thus helping to ensure that people continue to do what they do well and are supported by social media, rather than driven by it. In order for Twitter to be part of the disaster management team, it must help address the three key problems in disaster management are communication, coordination, and the exercise of authority (Quarantelli, 1988). On the one hand, the key is the provision of accurate, timely, relevant and geographically situated information (Guha-Sapir & Lechtat, 1986; van Oosterom, et al., 2006). However, the exact information needs will vary by the phase of the disaster. On the other hand, the key is understanding the biases in the incoming data so that it's accuracy and relevance can be taken into account, and so inform decisions. The socio-cultural-technical context will impact the accuracy, timeliness, relevance, and geographic features of the incoming data by effecting various biases. In this section, we identify these overarching biases.

2.1 - Twitter

Twitter is a 'microblogging' service, where users sign up for free, choose a 'handle,' and get a profile accessible at Twitter.com/[handle]. On these profiles, they can post messages ('tweets') of a maximum of 140 characters. By tagging other's user-handles with "@" symbols, they can have their tweets appear on the profiles of others, thereby engaging in conversations. By adding a hash symbol ('#') before a word, it is made machine-readable and called a 'hashtag.' These hashtags can serve multiple purposes, such as labels, summarizations, and topic indicators. By posting their own hashtags and searching for given hashtags (e.g., #worldcup),[2] users can participate in larger conversations, and get to know users to whom they are not currently connected. Users can also 'follow' other users, automatically seeing all of those users' tweets appear on a section of their profile.

Twitter's community and conventions have now been extensively studied (Honeycutt & Herring 2009; Marwick & Boyd 2011; Bruns & Moe 2013; Gaffney & Puschmann 2013; Rogers 2013; van Dijck 2013; Kumar, Morstatter & Liu 2015). This work recognizes that the technical affordances of Twitter are only one aspect; equally important is the type of community and norms that have arisen from Twitter. The twittersphere is often a vibrant place of discussion for social and political topics and a site of unique communities,[3] although much of its adoption (at least in the US) seems driven by people's interest following the Twitter accounts of celebrities (Hargittai & Litt, 2011). Because it gives a pulse of a particular segment of the public on certain topics, there are claims that following patterns on Twitter can help make accurate predictions about future trends, such as the box office success for a major movie on its opening weekend (Asur & Huberman, 2010; see also Wong, Sen & Chiang, 2012; Gayo-Avello, 2012). However, Twitter is not representative sample of any population; in the US, it is biased towards urban, affluent populations (Hecht & Stephens 2014; Malik et al 2015; Mislove et al. 2011). The demographics of Twitter users vary internationally (Poblete et al 2011), which affects our ability to make accurate inferences (Cohen & Ruths, 2013). Nevertheless, Twitter represents a vast and complex ecosystem, with corporate accounts trying to engage users, bots ('robot' accounts run automatically with computer scripts) that spam people (Donath, 2007; Thomas et al., 2013; Thomas et al 2011), users following celebrity intrigue, participating in political discussions, and carrying out harassment against others (Donath, 2007; Thomas et al., 2013; Thomas et al., 2011).

Since its launch in 2006, Twitter has grown to, as of 2014, 284 million monthly active users, who send about 500 million Tweets are per day, of which about 80% are from mobile devices. Like all major social media platforms, Twitter has an API (Application Programming Interface to which developers and researchers can submit text-based queries in pre-specified formats to which Twitter will return text-only data (instead of including images, or rendering these data in a graphical interface). Developers may use these data to make applications on top of Twitter, such

---

[2] Note that hashtags are not sensitive to capitalization, and cannot include spaces. Hence the confusion of some Twitter users at the hashtag #nowthatchersdead at the death of Margaret Thatcher, thinking that the singer Cher had died. http://www.salon.com/2013/04/08/internet_confused_by_nowthatchersdead_hashtag_on_Twitter/ last accessed – Feb. 2015. Furthermore, it is not always obvious what a hashtag for a given topic would be; during the Boston Marathon bombings, #prayforboston won out as the hashtag of solidarity adopted by increasing number of users over other early alternatives. ([CIT]) Conversely, it is not obvious what topic a given hashtag may represent, as—like in other cyberculture—they are often idiosyncratic abbreviations that are code for certain practices. For example, #tbt stands for "Throwback Thursday," where users post old pictures along with this hashtag in a public and shared performance of their individual nostalgia. Finally, hashtags originally sent for one purpose are sometimes usurped for other purposes.

[3] See, e.g., http://www.buzzfeed.com/jwherrman/weird-Twitter-the-oral-history

as customized interfaces or ways of interacting with Tweets, and researchers may use the text data to study patterns on Twitter (Gaffney & Puschmann 2013; Kumar, Morstatter & Liu, 2015). An example of such a platform is TweetTracker (Kumar et al., 2011; Kumar & Morstatter, 2011). Another such example is the Twitter Tsunami system (Landwehr, et al, this issue).

The ‘Streaming API’ will return 50 per second, or 3000 per minute with the maximum not to exceed 300 tweets for various parameters such as all tweets of a given user, tweets with a certain hashtag, tweets sent by users who have the geo-code enabled and which fall within the latitudes and longitudes that define a box of geographic coordinates of interest to the observer, etc. However, when the number of potential matches to a query exceeds the rate limits, the sample may not be random and distort trends (Morstatter et al. 2013). The ‘Sample API’ gives a consistent 1% sample of all of Twitter, but with no ability to focus on any particular topic, user, pattern, or other parameter. For a substantial cost, Twitter makes available a ‘firehose’ of the complete stream of public Tweets; not only does this cost money, but a company that has access to a stream must also have invested in infrastructure for taking in, storing, and sharing vast amounts of data in order to be able to make use of this stream. Twitter also shuts down or suspends accounts that they have determined to be bots or that violate Twitter policies (Thomas et al, 2011); which means historic data is often not collectible. The impact of this policy can be quite severe; e.g., in some cases up to 25% of the geo-tagged accounts in a country may be dropped (Wei & Carley, 2015). Further, users may have multiple accounts each with a different handle.

In Twitter, not all accounts are equal. There are ‘verified’ accounts, accounts that Twitter has investigated and determined that they do indeed belong to whatever public figure claims them. Being investigated for verification is based on a number of factors such as the extent that others want to find that user. High profile entertainment, news, corporate, and political accounts are often verified. This serves a vetting purpose, to let people know that the account may be taken as actually that of whichever figure or entity it claims to represent. This verification is also useful for research purposes, since these accounts are guaranteed not to be bots, but accounts with such a designation are a narrow sample of major celebrities and other public figures, corporations, and government agencies. Finally some accounts belong to individuals, some to corporations, some to news agencies, and some to government units. Different types of users, particularly governmental and corporate, may have their own policies on how they use Twitter, what hashtags are allowed, and whether they can refer to or retweet other users.

Since Twitter provides a means of getting vast amounts of linguistic, temporal, network, and other forms of data from actual human behavior, with relatively little effort, it has become enormously popular with computer science and social science researchers. Social scientists have shown that Twitter activity is a biased representation of humanity (Weller et al., 2014), with its own norms and practices rather than being a generalizable representation of human behavior and activity (Tufekci, 2014 Ruths & Pfeffer, 2014). It is important to characterize the nature of Twitter activity, and individually establish whether particular patterns in Twitter have external validity. For disaster management, among the unanswered questions is whether the data from Twitter will be useful for management and early warning, and if so how should it be collected to be of utility. In other words, even though the data is biased and not necessarily representative of the entire population, is the data sufficiently comprehensive, and with sufficient information on the spatio-temporal patterns and the users that it can be used to support planning and early warning?

2.2 - Twitter Usage in Indonesia

Indonesians were early adopters of Twitter and are among the most prolific Twitter users. In 2010, the comScore report ranked Indonesia as the country with the highest Twitter penetration, with 20.8% of the Internet-using population visiting Twitter in the month of June.[4] In early 2012 a market research firm[5] put the Indonesian Twitter-using population at 29.4 million, the fifth largest in the world in raw numbers. By 2013 CNN dubbed Indonesia the “Twitter nation” (Lim, 2013). By 2014, Indonesia was ranked as the fifth most tweeting country, Twitter reports approximately 29 million Indonesian users, and Jakarta was responsible for 2.4% of the 10.6 billion Twitter posts made between January and March of 2014.[6]

The 29.4 million Twitter users in 2012 represented only 11.9% of Indonesia’s total 2012 population of 247 million.[7] Still, this is impressive considering the World Bank estimated the Internet-using population of Indonesia at 15.4% that year,[8] meaning that 77.3% of Indonesian Internet users were using Twitter. In the same market analysis, out of the public tweets geo-localized at the city level, Jakarta was the most active city in the world, accounting for 2% of the total volume, and with Bandung coming in sixth. Another study (Poblete, Garcia, Mendoza & Jaimes, 2011) found that Indonesia has the highest tweets/user ratio at 1,813.53, ahead of Japan (1,617.35) and Brazil (1,370.27).

2.3 - Sociocultural Context

Between the migration of millions of young Indonesians from rural areas to the cities, and the rapid adoption of social media in these urban areas, Indonesia has witnessed major socio-cultural changes in the last two decades (Miguel et al, 2006). Although less than 1% of users had access to the internet in 1998 (Basuki 1998), it became a means to diffuse critical information. For example, when Suharto shut down other media, information spread through Twitter was critical to his being ousted. Today, the internet and social media have become strongly integrated aspects of the Indonesian socio-cultural landscape. Early on, Indonesian’s referred to Internet cafés, and sometimes the internet itself, as *warnet*, a contraction of ‘*warung* Internet,’ named after a simple place for lower-middle and lower classes to buy snacks or meals and congregate while eating (Lim, 2003). These warnets became a point of relatively inexpensive, and social, access to the internet outside of universities. Like pubs in England and America, people could go to the warnet to socialize, get access to news, and pass on news both through traditional social networks and internet based social networks, often outside the bounds of the official news channels. Thus, from the start, internet access was woven into the traditional family-friend networks of communication. The rapid adoption of social media, as more people moved to the cities and more people bought cellphones, occurred in such a way that the social media has pervaded all aspects of Indonesian life, and led to cultural changes and new socio-cultural behaviors. Examples include, cell phones being used to communicate with the world of the supernatural, such as a channel by which ghosts can communicate (Barendregt & Pertierra, 2008), evolving

---

4 http://www.comscore.com/Insights/Press_Releases/2010/8/Indonesia_Brazil_and_Venezuela_Lead_Global_Surge_in_Twitter_Usage last accessed – Feb. 2015.

5 http://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US last accessed – Feb. 2015.

6 https://www.techinasia.com/indonesia-social-jakarta-infographic/ last accessed – Feb. 2015.

7 http://data.worldbank.org/indicator/SP.POP.TOTL last accessed – Feb. 2015,

8 http://data.worldbank.org/indicator/IT.NET.USER.P2 last accessed – Feb. 2015.

images of women and public intimacy (Brenner, 1999), new internet-based economic concerns (Couldray & Curran, 2000; Lim, 2003), and public attention to fraud (Setiyono & McLeod, 2010; The Asia Foundation, 2014).

Internet and Twitter usage are concentrated in Jakarta and other urban centers in Java; however, the use of such technologies is generally relevant, especially in rural communities outside of Jakarta and Java. Much of the internet access is through mobile technology. By August 2013, 64% of the iPhone users in Indonesia, tweeted, as compared to 36% of the iPhone users in the United States.[9] Most of the Indonesia tweets appear to be about sports or local news. In 2011, a study of Asian tweets showed that approximately 20% of all tweets in Indonesia concerned soccer, 11% local events, 5% each for news, TV shows and music.[10] Among the most followed Twitter sites are @Indonesia, @IndonesianIdol, @BNI46 (Official Twitter of Bank Negara Indonesia), followed by the sites for various politicians and soccer stars. The top three of these sites being followed by people inside and outside of Indonesia. Overall, Twitter is not just integrated into the daily lives of Indonesians; its usage is from mobile devices and the bulk of the communications are directed at attending to sports, entertainment, and political-economy.

Information and communication technologies are widely used in disasters and their aftermath in Indonesia. During the 2004 tsunami in Aceh, community radio stations played a key role in disseminating information and broadcasting messages to help survivors recover from grief, after large parts of the media infrastructure were destroyed (Birowo, 2013). There was a similar use of community radio in West Sumatra when a 5.8 magnitude earthquake struck in March 2007. Community radio stations disseminated information from the Indonesian Meteorological and Geophysical Agency to reduce the impact of rumors that had begun circulating about an impending tsunami and another earthquake. The Community Radio Network of West Sumatra "dropped volunteers in the affected areas to set up emergency radio. Because their station budgets had no available funds, volunteers had to use their own money. They came from various community stations surrounding the affected areas. They alternated in working for 3 to 4 days at a time." Radio stations also broadcast *nazam*-s, traditional Acehnese poems, consisting of religious messages to help survivors recover from grief. Birowo (2013) emphasizes that this was a spontaneous response of Indonesians using available information and communication technology resources to spread information when other information infrastructure was destroyed.

As more Indonesians use social media, those specific technologies have become of greater interest to government officials and first responders in Indonesia as communication modes for disaster management. Anggunia and Kumaralalita (2014) give, from their perspective of an Indonesian provincial government analyst, a description of the use of Twitter and other social media in the lead-up to the eruption of Mount Kelud on February 13, 2014: government agencies posted warnings and geological updates on Twitter, and community groups and individuals helped spread information. After the eruption killed 16, the community radio helped spread messages of support, and provided key information.

Another system for testing evacuation routes following earthquakes based on tweets has been tested with respect to the Sumatra earthquake (Ishino et al., 2012). The key is that in major cities

---

[9] Sept 2013 - http://venturebeat.com/2013/09/26/Twitters-global-mobile-reach-36-in-u-s-8-in-germany-64-in-indonesia/ last accessed – Feb. 2015,

[10] https://wiki.smu.edu.sg/digitalmediaasia/img_auth.php/5/56/Trending_Topics_Twitter_Indonesia_2011_Saling_Silang.jpg last accessed – Feb. 2015.

there appear to be sufficient geo-tagged tweets to make route tracking feasible. While government officials turned to Twitter as one means of disaster management after the Mt Sinabung eruptions, the communication was not as effective as it could have been given that the failures in inter-agency coordination and the lack of clear cross-jurisdictional communication created challenges that the technology alone could not overcome (Chatfield et al., 2014; Chatfield et al., 2013). As of Feb 2015, not all of these difficulties have been overcome; yet, Twitter is currently being used as part of the disaster management process. Further, the Twitter stream appears cluttered with non-disaster information and extraction of relevant data requires topic modeling. New disaster-focused topic models show promise, and have been tested given Indonesian earthquake Twitter data (Kireyev et al., 2009). For early disaster warning and disaster management, Twitter can give lead time, but is often best used in Indonesia in conjunction with traditional warning systems and news (Chatfield and Brajawidagda, 2013).

The Australian University of Wollongong's SMART Infrastructure Facility and local Indonesian emergency agency, Jakarta's BPBD DKI, under a Twitter Data Grant, launched project Peta.Jakarta. This system employs a real-time mapping system that uses tweets to understand how floods are moving and what areas are hardest hit.[11] As another public safety initiative, Indonesian government agencies banded together to create a Twitter based early warning system for Tsunamis. An empirical examination of this system (Chatfield and Brajawidagda, 2013) revealed that key issues had to do with getting the information to the official users soon enough. This study did not consider how the public attended to this information. Chatfield and Brajawidagda (2013) note that in this system, the National Disaster Management Agency (NDMA) serves as the lead agency in national disasters and it works closely with the Meteorological, Climatological and Geophysical Agency of Indonesia (BMKG). The NDMA is the lead agency for tsunami detection and early disaster warnings and is responsible for operating the Indonesia Tsunami Early Warning System (Ina TEWS) to detect, analyze, simulate, and disseminate disaster information on earthquake, tsunami, and severe weather. This dissemination goes through multiple mediums including Twitter.

BMKG issues three types of warnings. The first early tsunami warning can be issued during the first 5-7 minutes after the occurrence of an earthquake. The second early tsunami warning contains either confirmation or cancellation of the first early tsunami warning and can be issued within 10-30 minutes. The third early tsunami warning is issued after the coastal observation of a tsunami and can be issued between 30-60 minutes. These early tsunami warnings are intended to inform government agencies; they are not issued directly to the public. However, BMKG is widely followed by the public on Twitter and it's warnings are retweeted (e.g. in 2012 on average BMKG warnings were retweeted 2000 times). We found that by 2012 BMKG had over 300,000 followers on Twitter, and today has over 1 million (see @BMKGIndonesia). Other key agencies that are part of the Indonesian tsunami early warning system are also widely followed. This includes: Polri e.g. @NMTCPolri, @BNPB_Indonesia, Kominfo, the Pacific Tsunami Warning Center, and @NewEarthquake.

2.4 - Disaster Context

There are many ways in which the kind of information shown to be available in social media can help in disasters. One of the best ways to help individuals recover from disasters is to empower them to work towards their own activity. Individuals act as rational actors, even in the

[11] https://gigaom.com/2014/12/04/indonesia-is-mapping-jakarta-floods-in-real-time-using-Twitter/ last accessed – Feb. 2015.

wake of disaster, and providing them with additional information can help them take steps to ensure their own well-being (Dynes, 1970). Social media supports this type of community resilience by enabling local actor to self-mobilize. Another way to help in disasters is to link responders to those in need. Social media supports this by creating a forum where requests for information or aid can be rapidly communicated. Whether uncovered on local blogs by relief managers or delivered by text message to the appropriate organization through intermediaries, this information is valuable (St. Denis et al., 2012; Munro, 2013). Another way to help is by providing early warning, or educational materials for how to prepare for the upcoming disasters. Social media supports this both because local actors can provide warning sometimes before sensors (as in the case of Volcanoes) and because this early warning and informational materials can be disseminated in a form that is easy to re-disseminate. Thus, getting the power of the crowd to work for you.

But which social media? To be sure, during a disaster or in response to it most communities employ multiple social media and the exact set employed depends on the community; e.g., Weibo in China, VK in Russia, Facebook Twitter and Instagram in Indonesia. However the research is dominated by assessments of, and tools for, Twitter. While Facebook set many of the standards for modern social networking sites, Twitter has become a tool of choice for researchers interested in studying responses to current events. There are three primary reasons for this: design, adoption, and data availability.

While Facebook set a standard for modern social networking platform design by only allowing users to post messages to contacts who had acknowledged them, Twitter adopted a model where users broadcast their messages to anyone who chooses to "follow" them – in other words, to treat the platform as a broadcast medium. The ways in which the news media, small relief organizations, the police, and other organizations have used Twitter to facilitate their public relations continues to be an important topic of research (Armstrong & Gao, 2010; St. Denis et al., 2011; Panagiotopoulos et al., 2011; TwitterMedia; Libby, 2015). The 140-character limit on tweets forces users to be brief, and makes it easy for anyone reading tweets to review many of them extremely rapidly. Further, there is a platform-wide convention of using hashtags—a hash mark (#) placed before a word—to associate a tweet with a particular topic or idea. This delineation makes it relatively straightforward to parse tweets for salient elements.

In 2012, Semiocast[12], a provider of real-time web consumer insight solutions, estimated that the population of Twitter users had grown to 500 million and that the number of tweets sent had reached 340 million per day (Semiocast, 2012). That same year, Alexa[13], a web analytics provider, ranked Twitter as the eighth most popular website in the world (Fitzgerald, 2012). It has largely been positively portrayed in the press, and is known for having played a role during the Arab Spring. While Twitter will not be around forever, it is currently a critical element of media consumption for many people, and is a defining part of our cultural zeitgeist.

Perhaps most significantly, Twitter has supported a policy of open data access, providing an API that allows anyone to relatively easily sample a stream of tweets using particular search terms. Consider this in contrast to Facebook, which generally requires users to provide consent before allowing access to their data. To use the latter platform to sample a stream of public opinions at a particular time requires a tremendous number of users to authorize access. On

[12] http://semiocast.com/en/ last accessed February 2015
http://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US last accessed February 2015
[13] http://www.alexa.com/ last accessed February 2015

Twitter, a 1% sample of tweets can be obtained at any time, and larger subsets are available for purchase.

Experiments comparing the 1% sample with the entire “firehose” of data have shown that the sample isn’t useful as a proxy for the true count of tweets about a particular subject (Morstatter et al., 2013). However, complete representativeness of all tweeters isn’t necessary when trying to support disaster response. What the “1% feed” provides is all of the tweets that match a particular set of search queries, up to 1% of the total volume of tweets passing through Twitter. The tighter the queries, the more likely you will get all the tweets. Many disasters are only of local interest, making it unlikely that the number of tweets using disaster related language will pass the 1% threshold.

Disasters that make national and international headlines, such as Typhoon Yolanda/Haiyan in 2013, will lose some volume of tweets. This is primarily a concern if the storm is producing a large number of headlines in advance of the event, boosting awareness and pre-emptively dominating the conversation. While this can’t be completely solved through technical means, if a response agency lays an initial groundwork of promoted hashtags and dedicated accounts to which victims can send information, they can limit the effects of saturation.

The ease of acquisition, volume of data, and crowd-sourced nature have led many researchers to build tools to leverage the data produced by Twitter's users and to use the data for a variety of predictive ends, such as predicting film revenues, modeling the path of disease outbreaks, and locating the epicenters of earthquakes (Asur & Huberman, 2010; Sadilek et al, 2013; Earle et al., 2010). These applications tend to focus on looking at the rates at which and locations from where users post particular phrases. This combination of elements has also drawn a large number of organizations, first responders and government organizations to explore the use of Twitter for disaster response. Other research has focused more on the human aspects of Twitter, such as how individuals use Twitter during disasters, or the network structures of Twitter communication after disaster events (Acar & Muraki, 2011; Sutton, 2010; Mendoza, 2010). Another thread has focused on proposing particular ways of structuring tweets to make communication more efficient. What none of these applications has done, is to use Twitter to provide guidance on the dynamic flow of the population throughout the day, to estimate which part of the population can be reached electronically, or to identify local opinion leaders to help mobilize the public. Our system, the Twitter Warning and Response Social Media System (TWRsms) addresses these issues.

As previously noted, Twitter has a potential to be used in all phases of disaster management and response as part of the overarching knowledge management system. Our specific concern herein is with planning and early warning, and to a lesser extent, needs identification and management after an event. While knowledge management is not a new concept for disaster responders, each disaster is generally treated as an entirely new problem for which responders build new knowledge structures on the fly and decision makers take action on the basis of this hastily constructed data (Yates & Paquette, 2011). Despite good intentions, poor decisions often result from inaccurate information. Information that is comprehensive, consistent, and accurate supports improved planning and decision making. A knowledge management system for disaster management needs a number of capabilities including (Zhang et al., 2002):

- Ability to predict the general nature or trend of the disaster
- Ability to evaluate the severity of the disaster, by population and location
- Ability to generate timely and specific warnings, that can reach the populations at risk

From this perspective, there are a number of questions that need to be addressed to understand the utility of the Twitter data. We focus on the following three questions: coverage, spatio-temporal patterns, and opinion leaders.

Coverage: How "good" is the Twitter coverage of the local population? As we have seen, the timeliness and crowd-sourced nature of Twitter, and it's use by the authoritative sources such as news providers and BMKG, and its widespread use, all point to the potential of Twitter as a key feature in any knowledge system for disaster management. However, as was also mentioned, the population on Twitter may not be socio-demographically representative of the underlying population. Understanding who is using Twitter is important for a number of reasons. If the volume of usage is too low, then it may not be worth investing scarce resources into Twitter support for disaster management. The composition of those using Twitter will provide guidance on which sub-populations are potentially reachable by this media during a disaster, and which will need to be reached by other means. Similarly, the composition of the user community along with the volume will be needed to determine what other data will be needed for planning so that the entire population is covered. The language used in the Tweets will provide guidance for the language in which early warning alerts should be sent. And so on. The danger of using internet-based technologies that are not held representatively is that decisions based on these may reinforce existing inequalities or create new ones (Thomas, 1995), lead to the propagation of rumors, and provide an illusion of accurate response (Quarantelli, 1997). As is note by Mackay (1995, p. 47) "The use and meaning of information and communication technologies in the home [and the workplace] … can only be understood within the class, gendered, geographical and generational context of its consumption." Thus, while Twitter data is potentially valuable, its accurate interpretation will require establishing a baseline for its interpretation and understanding the inherent biases in the data due to either the way it is collected or by whom it is contributed. Understanding the biases in the coverage is critical for determining how to use Twitter as part of a disaster management strategy.

Spatio-Temporal Patterns: How "good" is the Twitter coverage of where the population is when? Socio-demographics, however, are not the only issue. For example, geo-information provides the opportunity to aid in evacuation and response; however, the value of such information depends on its accurateness and up-to-date-ness (van Oosterom, et al., 2006; Earle, *et al.*, 2011). Social media can be used to map crises, see for example the 'first ever official United Nations crisis map entirely based on data collected from social media" (Meier, 2012). Tweets can contain overt geographic information in two ways – either the user can elect to share their latitude and longitude, or they can list a location – e.g., Padang. The latitude and longitude information has been used to support disaster management and response activities (Power et al., 2014). However, the overall geography of Twitter is more complex and can be better discovered also taking into account the location and language information (Leetaru et al., 2013). Still, we have a tendency to use data if it can be measured; and if that measurement is flawed, then there is a higher likelihood that the decision is flawed. In the case of Twitter relatively few tweets have any geo-information; thus, the critical question for disaster management is how geographically representative is the user population. That is, are the people where the tweets are?

The temporality of the data is an issue for disaster management. On the one hand, q key feature of many disasters is that power either goes out or becomes intermittent, thus mitigating the use of computer-based technologies during such times (Quarantelli, 1997). However, there is a more pernicious problem that impacts response and early warning; specifically, the technology may not be being used 24/7. Pattern of life activity with respect to a media will be a determinant of the extent to which that media can be relied on to gather data from, make projections based on, and alert the population. Since these patterns vary by location, the full spatio-temporal pattern must be assessed. Understanding the spatio-temporal patterns in Twitter usage is critical for determining the sufficiency of Twitter for signaling where people are, and so supporting evacuation and response efforts.

Opinion Leaders: Can Twitter be used to identify the local opinion leaders, and so mobilize them to help provide information to the populace? One of the unique features of Twitter is that it can be used as a news source and so support information diffusion. As noted, in Indonesia, authoritative sources such as BMKG are currently using Twitter to disseminate information. However, there is another source of authority in the Twitter-sphere and that is socio-cultural authority. This is the authority that a user attains through being an opinion leader – e.g., by having a large following, being highly central in the co-mentions network, frequently sending tweets that are retweeted and so on. Understanding who these actors are, and whether they can be systematically identified and mobilized, is critical to harnessing Twitter for early alerts.

3 – Data Collection

If Twitter data is to be used for disaster management it is important that it be collected in a systematic fashion and that the biases in that data be well understood. In this section we ask, what are the strengths and limitations of different data collection strategies as they relate to issues of coverage, spatio-temporal patterns, and identification of opinion leaders.

As noted, unless the full firehose is purchased, the data stream used in a disaster management tool will be a sample. From a disaster management perspective the goal would be to collect those tweets coming from within the region of interest, by the population in that region, with a principle focus on issues related to disaster management (in other words, minimize the number of tweets collected about general politics, sports and entertainment). However, collecting tweets only from a region of interest by the true population of that region and only on the topic of interest, from the Twitter API is difficult. For example, there is a parameter for "country" in the tweet metadata, but it is user-reported rather than tagged by Twitter, and furthermore it is not a parameter by which one can make requests to the API; however it can be used to post-process the tweets and remove unwanted tweets from the data collection. As such, the data collection strategy will impact the coverage, geo-temporal pattern, and ability to identify opinion leaders.

We considered five different collection strategies (CS):

- CS1 - Decahose Indonesia: a 10% decahose from which only tweets geo-tagged to Indonesia are used
- CS2 - Decahose Padang: a 10% decahose from which only tweets geo-tagged to Padang are used
- CS3 - Streaming Sample geo+terms Indonesia: 1% sample from the streaming API that uses a bounding-box on Indonesia and selected disaster related key words,
- CS4 - Streaming Sample geo+terms Padang: a sample from the streaming API that uses a bounding-box on Padang and selected disaster related key words,
- CS5 - Streaming Sample geo Padang a sample from the streaming API that uses only a bounding-box around Padang.

The two streams from the Twitter decahose (CS1 & CS2), are based on a generic 10% sample (worldwide) from which we post-facto selected only those tweets that had actual geo-tags (latitudes and longitudes) that were within either Indonesia, or just in Padang. The overarching archival sample was made available to researchers at our institution and contains over 100 million tweets (O'Connor et al., 2010; Einstein et al., 2014). The difference between CS2 and CS5 is when geo-tagging is used. For CS2, it is a 10% random draw where post-facto only those with geocordinates in Padang are used. For CS5, all tweets from a bounding box that includes Padang such that the total does not exceed 1% world wide are selected. This guarantees that they all have geocordinates. Then if there happen to be any outside of Padang they are removed in the same was as with the decahose. The result as is seen in Table 1 is more tweets in Padang.

All of CS3, CS4 and CS5 are collected using the streaming API which returns those tweets matching any of the criteria in a set of criteria, up to a total of 1% of all tweets in Twitter at that time. The full details of the infrastructure and criteria used to gather the samples from the streaming API is described in Landwehr et al. (forthcoming – this issue). The approach used is replicable using other infrastructure. Three different data streams were captured by three different users – one of which is a general Indonesia stream and two of which are more specific Padang streams. In all cases, relevant geo-bounding-boxes were used to set the area of interest. The size of the bounding-box was set to the minimum possible that would enclose the entire region of interest. In two of the streams a set of terms in English and Bahasa (Indonesian) related to tsunamic disasters were used to identify tweets of interest, in the other two only the bounding-box was used. We searched first for any tweets either falling in a bounding-box covering Indonesia (southeast corner -9.5, 95.0 and northwest corner 6.0, 141.0), or in a bounding-box covering Padang (100.25, -1.05; 100.5, -0.75). For CS3 &CS4, tweets were selected using both the bounding box and a set of terms containing one or more words from a list compiled by the researchers (in consultation with Indonesian colleagues) that includes the names of 53 Indonesian locations, critical Indonesian Slang words used with respect to disasters, the names of prominent Indonesian organizations and citizens, and disaster-related Indonesian words. To the extent possible, all words or phrases were in English and Indonesian. This will bias the sample towards tweets about locations, politics, and disasters, which is acceptable for disaster management.

Each of these strategies has pro's and con's. The purely random samples provided by the decahose better represents Twitter's population as a whole (CS1 & CS2). They can be useful for understanding how any particular facet of Twitter – presence in a country, use of language, particular hashtag- exists relative to the rest of the Twittersphere. It isn't particularly useful for mining detailed facts about a specific incident because it hasn't been targeted from the get-go. A streaming, searchable decahose sample would certainly be useful to first responders, but a streaming, searchable 1% is more useful as it focuses attention on the issues of concern. Note, while the scale and media coverage of a disaster can increase its mentions on social media, as a general rule few disasters will reach that 1% limit. In most disaster scenarios, the streaming API can be sufficient for monitoring Twitter as a whole.

The Twitter API, circa 2012-2015, places a number of restrictions on collection that further bias the data. It is possible, using the API, to request tweets with a geolocation that falls within the border of Indonesia using geo-spatial bounding-boxes. There are two difficulties here. First, less than 10% of all tweets are geo-tagged (Morstatter et al., 2013) with latitudes and longitudes. Thus, bounding-box selection only will bias the sample towards users unconcerned with broadcasting their specific location and using smartphones with GPS. Such Twitter users may be systematically different from others. Second, the Twitter API provides tweets from predefined polygons such that even when a bounding-box is the only selection criterion, tweets with geo-tags that fall in overlapping polygons, but not inside the bounding-box may appear in the collected data stream. Limiting the bounding-box to lie within Indonesia entirely, or using multiple tiny bounding-boxes that track Indonesia can mitigate this problem but it is non-trivial to set up as a collection strategy. We note that, even if not representative, geo-tagged tweets are extremely valuable because they provide precise location information. This is critical in disaster situations where the areas affected are far more specific than an entire country, and a geo-tag would give coordinates precise enough to direct first responders to, e.g., a damaged or collapsing building.

To request tweets falling within a geographic region a box or set of boxes must be specified (not a complicated polygon, as borders of countries are). Then it is necessary to filter the returned tweets to include only those falling within the region of interest; e.g., setting a box to cover Indonesia and then excluding those tweets falling within Malaysia and other countries with land in the bounding-box or with overlapping polygons. Finding whether a point falls inside or outside of a complex polygon is a mathematically complex and computationally difficult problem, but fortunately one that has been studied extensively, and for which the best solution is easily available to implement. We carried out such a filtering using existing tools for CS5.

When the data were collected using a bounding-box or terms of interest composite query, there were *zero* geo-tagged tweets returned for Indonesia (CS3) during some periods, and only a small handful for Padang (CS4). Though highly unusual, it is not impossible particularly for that collection period that coincided with the 2014 Iquique earthquake in Chile. This earthquake, and the resulting tsunami, were the topic of a great deal of Twitter discussion in Indonesia (see Landwehr et al., forthcoming). During this period, the keyword hits appear to have drowned out the bounding-box hits. However, there is an app called "UberTwitter" (http://www.uberTwitter.com/) which, for the users who use it, has a service to set the user-reported "location" field (distinct from the "country location" field. This is a field where many users exhibit a great deal of creativity rather than reliably saying where they actually are) to their current geocoordinates prefaced by "ÜT:". Thus, we used character matching to extract all such cases from our data set. As it turns out, 0.71% of tweets have this UberTwitter tag and accompanying geocoordinates, which is considerably less than the 3.17% of tweets from the Streaming API and 1.45% of tweets from the firehose reported by Morstatter et al. (2013) but still not insignificant. We hereafter consider the UberTwitter coordinates as the geo-tags. While UberTwitter users are likely yet another subset of users who are different in systematic ways (and using UberTwitter is one effect of this systematic difference), the idea that the tweets for which we have precise geocoordinates are a minority that are systematically different is the same for Twitter's geo-tags and for those of UberTwitter. Hence our method of investigating systematic differences and the nature of given subsets applies even if actual geo-tagged tweets would have their own differences.

Collecting tweets by language, is not simple. For the Twitter API , 'language' is not a parameter by which one can make requests, even though the tweet metadata contains both a field for the tweet language and for the language of the user's profile (both of which users must choose, and this information is included with every tweet). We do not collect by this meta-data term; but, as with location, we use it in analysis. Collection by language would exclude Indonesians who tweet in English, and include members of the Indonesian diaspora who tweet in Indonesian. Further, although Indonesian/Malay (specifically Bahasa Indonesian/Bahasa Malay) is the official language, over 700 languages are spoken in Indonesia.[14] Within large population

[14] Note Bahasa Malay and Bahasa Indonesian are the same language, and the official language of both Malaysia and Indonesia.

centers, many people speak English and tweet in English; and, English is spreading rapidly in part as English language education is compulsory in the schools. In the large population centers, where Twitter is predominantly used, the under 40 population who are the most common users are also the likely to speak and tweet in English. Thus only selecting tweets in Indonesian is likely to miss many important tweets by this community. With this consideration in mind, when tsunamic related terms were used as part of the selection criteria (CS3 & CS4) we used the same set of terms in both English and Bahasa so as to pull from the entire community in Indonesia. We defined a set of terms of interest based on prior Twitter studies of natural disasters with a focus on needs – food, water, shelter, safety, lost children or parents, and corruption. We then filtered out, those tweets associated only with celebrities or sports. We use the profile language and location in the meta-data as attributes of the tweet for analysis.

Working with Twitter data, it is also important to recognize that there are other possible biases that occur during collection. Morstatter et al. (2013) have shown that there are serious concerns about the reliability of the Twitter Search API, which returns a non-representative sample of high-volume data. However, considering that this is the source that would be available for monitoring in disaster situations, it is an appropriate stream to describe and analyze.

The consequences of these data collection strategies for the number of tweets and users and the nature of the tweets is shown in Table 1. Clearly, any form of collecting can result in a prodigious amount of data. Note, however, that the average number of tweets and users per day is quite different. So to is the dominant day and time period. While this summary indicates that the coverage is better using the API, it is not clear whether the inclusion of terms or not is better.

The key finding across data collection strategies is this: While access to the decahose may provide more tweets overall, the number of those tweets that are of specific interest for a specific disaster is low. Hence, selective collection even using the 1% API is likely to identify more tweets of interest. For disaster management, initial collection by the region (bounding box) has the advantage that it provides the largest corpus of data from the region of interest; however it will be biased as it will not include tweets from the region that do not include a geo-tag. For small regions, such a strategy may get you near perfect coverage of the geo-tagged tweets. The second point with respect to data collection is that the overall profile for how twitter is used may appear quite different under the collection regimes.

Table 1. Differences by Data Collection

Metric	**CS1 Decahose Indonesia**	**CS2 Decahose Padang**	**CS3 API Indonesia Box Terms**	**CS3.1 Sample**	**CS4 API Padang Box Terms**	**CS5 API Padang Box**
Start Time	1/1/2011	1/1/2011	2/24/2014	6/1/2-14	2/24/2014	11/05/2014
End Time	9/28/2014	9/28/2014	6/16/2014	6/7/2014	6/16/2014	On-Going
Tweets Collected	66,400,000	431,971	678,640,600	13,670,165	12,187,189	52,286,500
Day of week	Sunday	Sunday	Sunday	Sunday	Tuesday	Wednesday

with most tweets on average						
Hour of day with most tweets on average	8:00 PM	9:00 PM	5:00 PM	2:00 PM	9:00 AM	9:00 PM
Ability to find networks of mentions or retweets	Low	Low	Medium	Medium	Medium	Medium
Dominant Topics	Trending topics – sports, entertainment	Trending topics – sports, entertainment	Term related topics	Term related topics	Term related topics	Locally trending topics – sports, entertainment
Regional coverage	Proportional to Twitter usage	Proportional to Twitter usage	Driven by topics with small regional focus	Driven by topics with small regional focus	Driven by topics with small regional focus	Dominated by region of interest

3.1 – Current Results

In this section, we review the data that we've collected so far, noting two particular spikes in our data. We also look at how the keywords we collect are impacted by outside events, and consider the difficulty of making a useful filter. Because our system is currently being actively developed, with a close with a secondary example

We began collecting data from Twitter early in 2014. During this process we have modified both the search queries that we use to collect tweets and the management steps that we use to remove tweets and changed the parameters by which we cull tweets from the initial collection. We also lost several days of data from our Indonesia collection due to disk issues; regrettably, this includes the majority of tweets collected between March 8 and the start of April. The Padang collection was unaffected by these problems. Taking these factors into account, we collected 1,883,426 and 19,798,545 tweets posted during the months of March and April from our Padang and Indonesia collections respectively.

As can be seen in Figure 8, the data forms a sequence of spikes. The biggest of these, starting on March 7, 2014, is largely due to the disappearance of Malaysia Flight 370. The surge near the start of April is due to the April 1 Earthquake off the coast of Chile, which briefly posed a risk of creating a tsunami that would strike Chile. Correspondingly, users began to use "tsunami" more often, an effect that we can see captured in the percentage of our tweets that contain the word. (See Figure 9.)

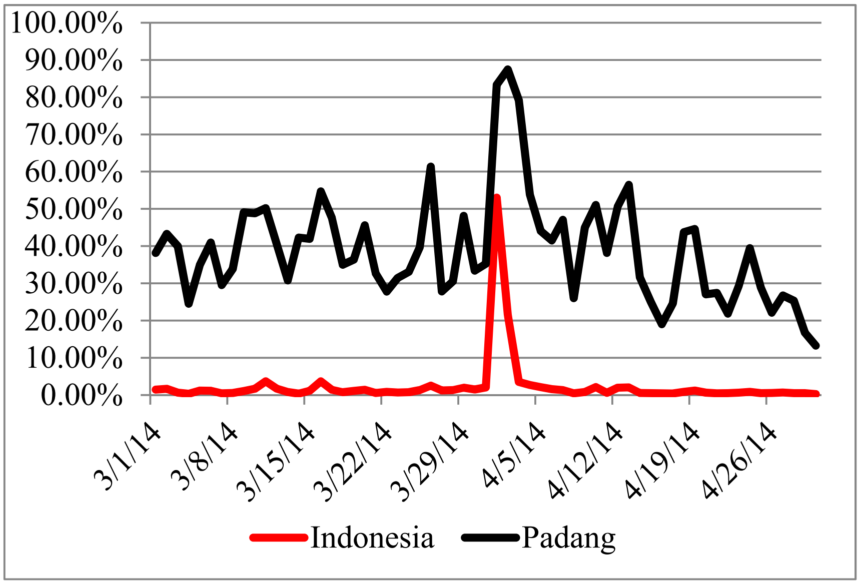

Figure 8: Total tweets collected from Indonesia and Padang during March and April, 2014.

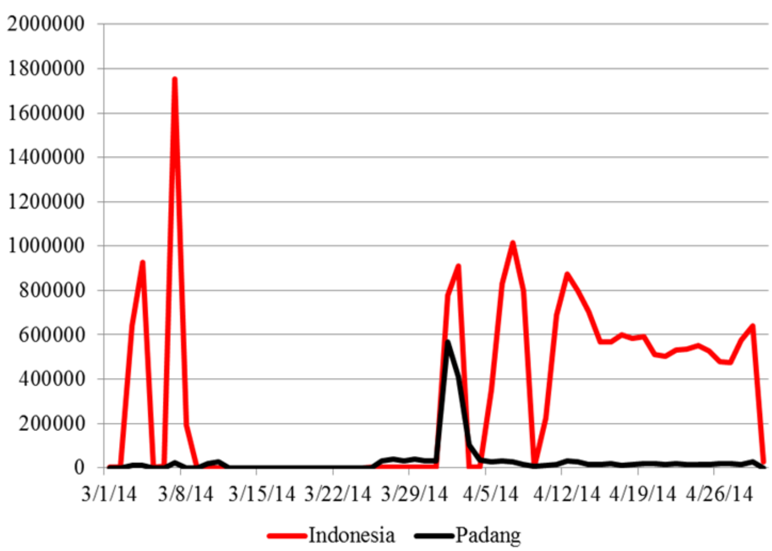

Figure 9: The percentage of tweets collected each day that use the word "tsunami"

As noted there was an unavoidable data loss. Such data losses are to be expected in such systems – particularly if they are to be run on devices in areas subject to power losses. A consequence of such a data-loss is that it's difficult to estimate the true size of the spike in the Indonesia data caused by the earthquake, but the representation in Figure 9 is dramatic. While the increased use of the word has potential as an alarm system, because of the constraints on the streaming API we aren't able to be certain about where on Earth the mentions of tsunami are occurring. Again, this constraint is that less than 10% of the tweets are geotagged. Additional filtering can be used to help guarantee that the information of relevance is from the area of concern, but it is unlikely to be perfect. (See Figures 10-12.)

Some of the April 1 tweets about the earthquake are indeed from Indonesia in general and Padang in particular. In Figure 10, those tweets with geotags for the area are shown. One implication of this is that these keywords are indeed picking up disaster related information. A second implication, however, is that much of the activity in Twitter when an event occurs is from outside observers, and not from those who will be or are being impacted by the tweets.

These spikes impact not just the number of tweets, but the network of tweeters. In Figure 11, the retweet networks immediately before and after the Chilean earthquake are shown. On the left (before) we see a sparse retweet network. This network show the classic starburst form of Twitter retweets. These starbursts are caused by Twitter assigning all retweets as connections to the original tweet and not to the tweet that the retweet actually got the retweet from. On the right (after) we see a larger more dense network. This latter network reflects the upsurge in concern, surprise, and response to the Chilean earthquake. The size and density of these retweets networks are different. Moreover, who is critical (for example who is most retweeted) changes.

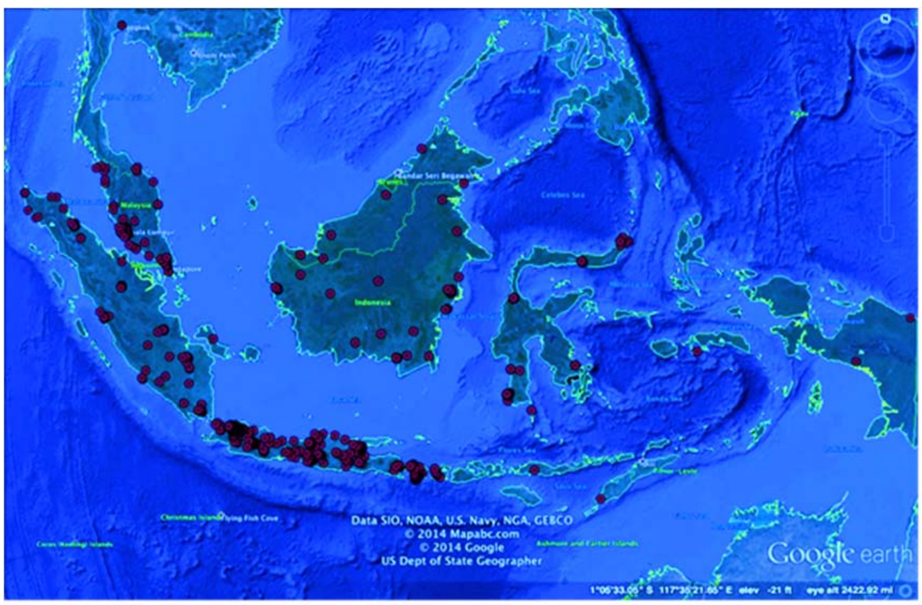

Figure 10: Map of all geolocated tweets we collected from Indonesia on April 1.

System flooding due to overuse of keywords, is a critical issue for systems such as ours. Some keywords can result in too broad of a data collection which obscures the signal of interest. However, we do not want to miss any important tweets. Dropping the search term risks losing valuable content. Further, this collection method doesn't address the problem of capturing useful tweets that don't use any of the keywords for which we have been searching. There are several ways to combat this problem. One would be to change our static collection method for one that is more adaptive. We described this approach in the Collection portion of our Methodology: we could modify the search terms and users that we are using to get tweets from the stream based on the content of the tweets that we have been analyzing. For example, we could start building a new set of terms by searching for words that often occur in the same tweet as "tsunami", since other tweets in which they appear may be tsunami-related. If so, we might then experiment by dropping tsunami as a search term to try and avoid the extra noise that it is bringing in while still getting the useful tweets that include the co-occurring terms.

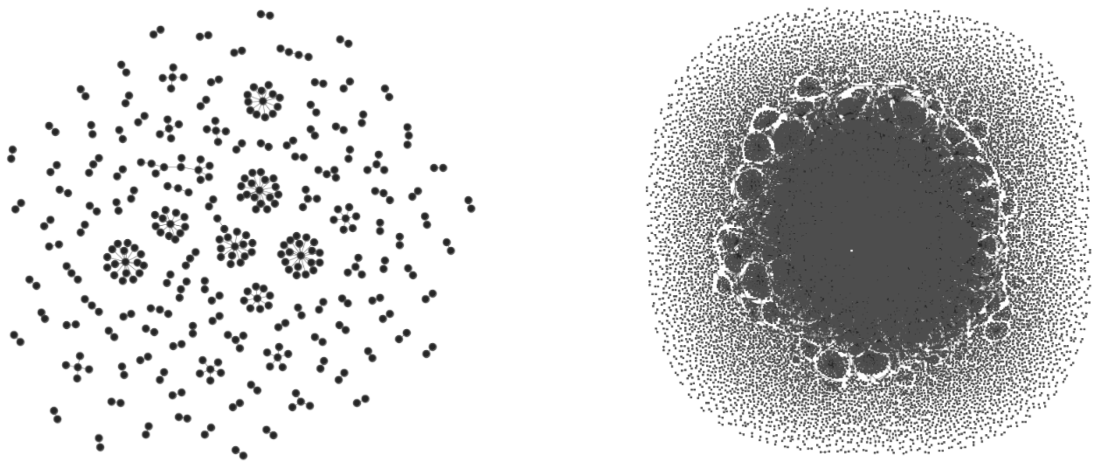

Figure 11: The retweet network from 11 PM on March 31 (left) prior to the Chilean earthquake and the retweet network from 11 PM on April 1 (right) after the Chilean earthquake.

Another method of culling captured tweets is to classify them on-the-fly as relevant or irrelevant to the disaster and then discard or retain them accordingly. The set of disaster-related n-grams that we described in the Introduction are one way to approach this problem; tweets without any disaster-related N-grams can likely be safely discarded. Figure 12 graphs the percentages of collected tweets that would be retained if we were to use such a system right now, an average of 58.52% for Indonesia and 74.07% of tweets for Padang. These codes are still being developed and refined, but a filter of this type is simple to apply and can produce meaningful reductions.

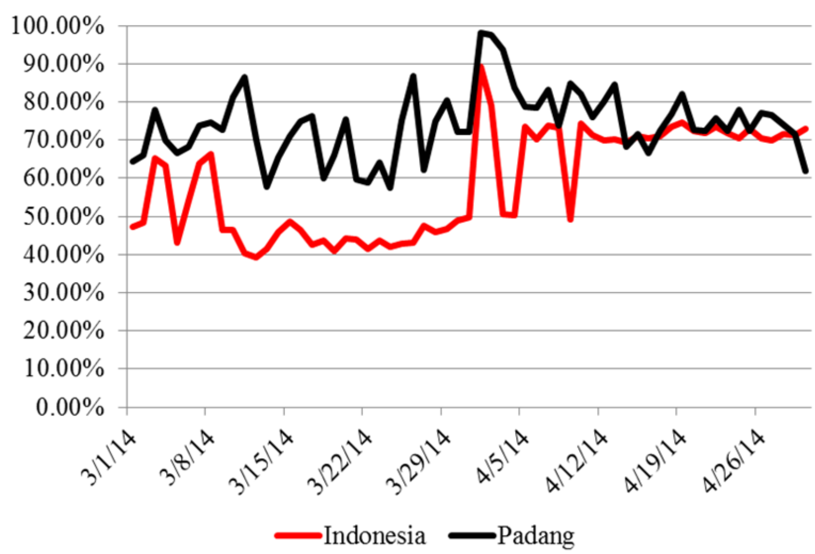

Figure 12: The daily percentage of tweets in each collection that we would retain if removed tweets with no disaster-related n-grams

The classification of tweets by disaster related terms provides a simple barometer for awareness of disaster related issues. Our analysis of these terms reveals that harm (H) and sheltering (S) tend to move together. Solicitations for aid (A) and general disaster relief (R) also tend to track each other; however, not as extremely as harm and sheltering. These findings suggest that our core disaster related terms still need refinement.

Figures 13 and 14 show the percentage of tweets assigned to each category for the two different collections. Even when there is no tsunami, we see an average of 13.44% tweets being assigned to each disaster-related category for both the Indonesia and Padang collections. This suggests some continual awareness. The exact pattern differs between Indonesia and Padang. In Indonesia in general, most of the disaster related tweets are general solicitations of aid, then concern with the damaged infrastructure, then expressions of harm, and only a few are concerned with crime.

In Figure 13 (Indonesia) and Figure 14 (Padang) we show the trends for general issues. In this data from Indonesia, at this time, there is little signal for or discussion of corruption and fraud. However, we note that in other situations such as Haiti, corruption and fraud were frequently discussed. Finally, concern with crime tends to increase when concern with damages increase. In Padang, Figure 14, there is more concern with disaster related issues. Expressions of concern with damage outweigh general solicitations for aid. Crime and corruption still show up low; however, there is relatively more concern with crime in Padang than in Indonesia more broadly.

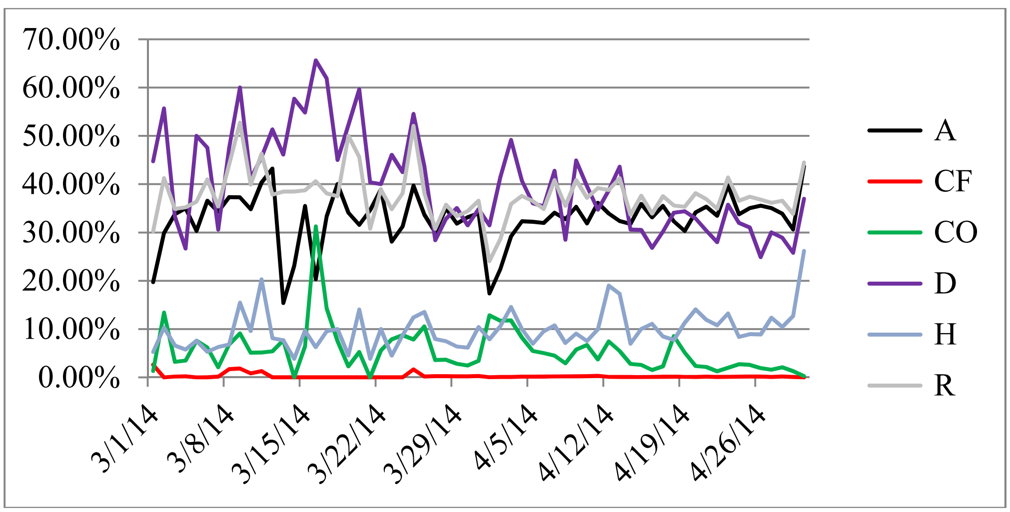

Figure 13: The daily percentage of tweets collected from Indonesia that fall into the general disaster categories.

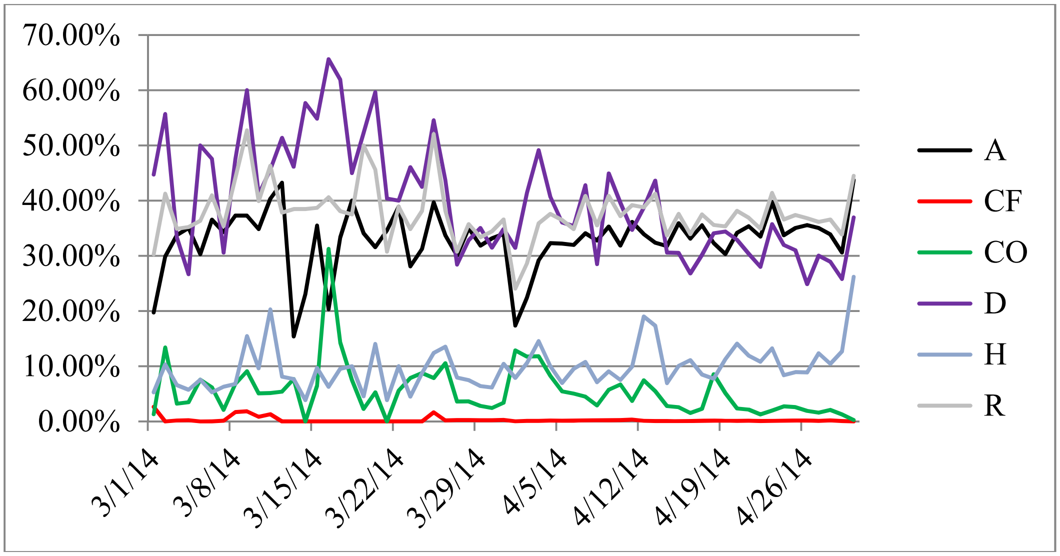

Figure 14: The daily percentage of tweets collected from Padang that fall into the general disaster categories.

If we consider the basic Maslowian needs, food/water, shelter, safety (missing people and fatalities). In Figure 15 (Indonesia) and Figure 16 (Padang) we see the signals from these terms. In Indonesia and to a lesser extent in Padang, concern with fatalities tends to be less than but correlated with, concern with damages. In Indonesia in general we see a greater concern with missing people prior to the April 9 election and a greater concern with shelter after the election. In general there is little concern with food and water, which is not surprising as at this point Indonesia was not facing a specific natural disaster. In Padang, although there is more concern with damages in Padang than in Indonesia more generally, there is about the same concern with the core issues. In Padang, the election did not cause as significant a switch in topics discussed.

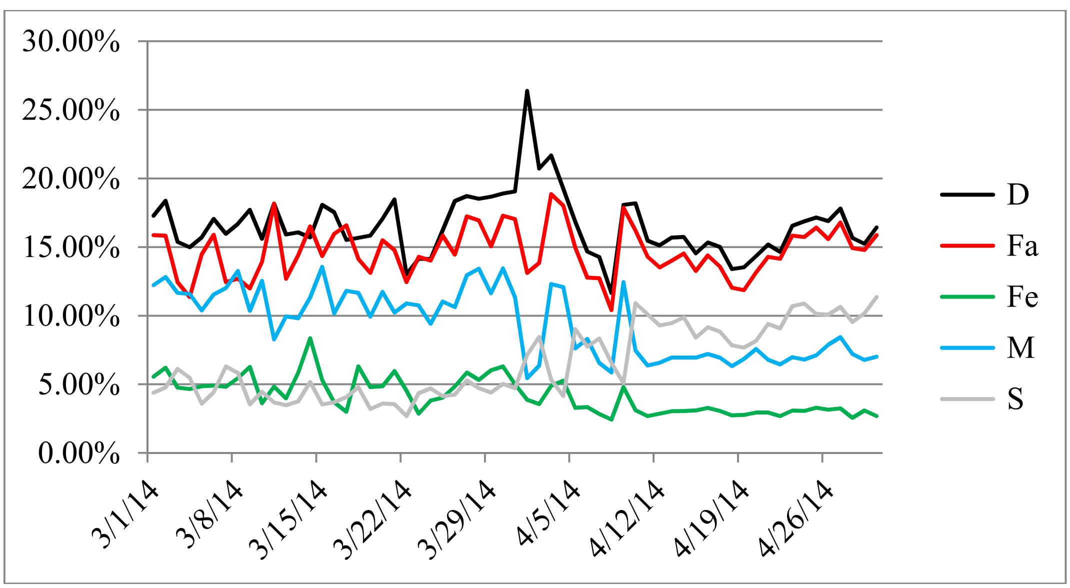

Figure 15: The daily percentage of tweets collected from Indonesia that fall into the basic needs categories.

From a crisis perspective, these results illustrate the following issues. Information about needs are generally a rare occurrence in Twitter. Thus strong techniques for identifying low volume signals are needed. Second, there is high variability in the extent to which discussions occur about the basic Maslowian needs, whether or not one is in a disaster. Third, extracting a signal about these needs is tricky and requires a strong basis of hand coded data, in each language of interest, which can serve as the gold standard for training the machine learning algorithms needed for classifying tweets.

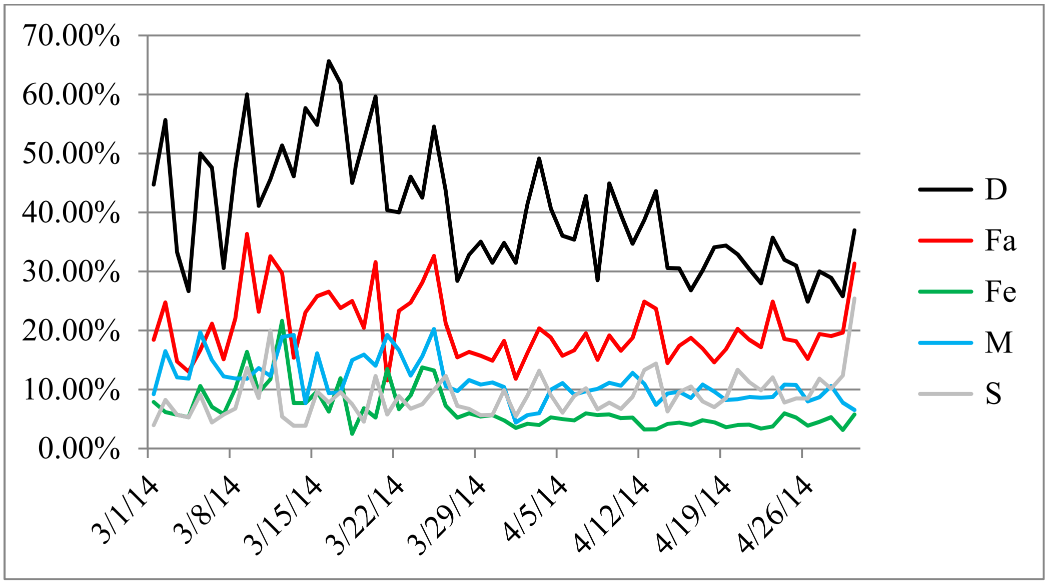

Figure 16: The daily percentage of tweets collected from Padang that fall into the basic needs categories.

4 – Tweets for Disaster Management.

We now examine the biases in the data, for Indonesia in particular, with respect to coverage, spatio-temporal patterns, and the identification of local opinion leaders. We note that the nature of the bias is a reflection of the way in which Twitter is used. Thus, in examining these issues we describe results using CS3; however, the reader should note, however, that a similar issue will arise regardless of the collection strategy.

4.1 Coverage

To what extent are the collected tweets representative of the underlying population; i.e., how good is the coverage? We begin by examining three aspects of coverage: volume, usage and language. As we just saw, the API provides better coverage in terms of volume. At issue is whether the inclusion of terms along with the bounding box distorts the coverage, and whether those covered are actually from the population of concern. The Twitter Streaming API lets the user define a set of search criteria. These criteria are treated as an OR, not an AND. Thus, it is possible that collecting by region and terms as was done in CS3 and CS4, may, in the case of a disaster reduce local coverage. For disasters, those outside the region are prone to tweet about the event; e.g., when the Chilean earthquake came most of the tweets in our two API collections using terms appear to be outside Indonesia and Chile and simply commenting on the disaster.

We now take a systematic look at coverage by focusing on the data for CS3. This period was chosen as it is a fairly normal period (no major external disasters), we had full coverage (no server outages or Twitter breakages), and it is a sufficiently large set that general trends can be assessed.

4.1.1 - Volume

As previously noted, there are multiple types of users. In Table 2, the volume associated with three classes of users is shown. Note, that News and Verified Users are tweeting at a higher rate and with less variability than are others. There are three key implications for disaster management. First, the news as an authoritative voice is well and consistently represented. Second, there is a great deal of bias in this data to issues not related to disaster or to Indonesia in particular as signaled by the high involvement of the verified actors and the large number of languages and locations. Third, the number of verified users, and tweets from these verified users, is extremely small.

Table 2. Summary Statistics for Indonesia

	Tweets		Users		Tweets per User	
	Per Day	Per Month	Per Day	Per Month	Per Day	Per Month
Overall						
Average	2,143,260	9,644,672	12,856	51,459	167	187
Std.Dev.	(4,054,835)	(8,944,601)	(10)	(12,235)		
News						
Average	1,509	6,792	6	23	252	295
Std.Dev.	(4,106)	(6,556)	(10)	(18)		
Verified						
Average	6,999	31,496	12	68	583	463
Std.Dev.	(16,130)	(31,590)	(32)	(63)		

While verified status is helpful in identifying sources that we may take as having higher content credibility, verified users and their tweets are too sparse to be useful when used alone for looking at patterns in volume. This is especially true when we look at intersections of tweets from verified users and other subsets: among all geo-tagged tweets, there are 184 tweets from 10 verified users. Among tweets with coordinates in Indonesia, there are only 179 tweets from 8 verified users (@AzmiShabana, @CloserOnline, @IndosatMania, @OfficialRCTI, @ToyotaID, @alfamartku, @barelybrad, and @radityadika). Although the tweets with user-reported country location are otherwise greater than those with an UberTwitter geo-tag, among tweets with any user-reported country location, there are even fewer tweets and users at 115 tweets from 13 users, and among tweets with Indonesia as the user-reported country location, there are 86 tweets from only 2 users (@BANKBRI_ID and @NatashaSkinCare). Thus for disaster management, the focus must extend beyond verified users.

Next we consider the volume of tweets by language, user's purported language, and location – see Table 3. These are CS3.1 data. Not surprisingly, the majority of tweets are in Indonesian or English, and by users tweeting in those languages and saying their profile that they use that language. Note that even though most of the users in this sample list English as their language, most are tweeting in Indonesian. Although not shown in the table, we note that, only .03% of these users have multiple languages in their profile, however, there are a substantial number of users who tweet in more than one langue – 330,040 (9.96%). Of these tweets, very few have either geo-tags or locations mentioned. However, where the location is indicated most are in Indonesia.

Table 3. Overall volume of tweets and users for Indonesia

Subset	Tweets	Users
All tweets	13,670,165	3,312,550
Tweets in Indonesian*	8,530,352 (62.40% of all tweets)	1,515,698 (45.76% of all users)
Tweets in English*	4,272,018 (31.25% of all tweets)	1,732,130 (52.29% of all users)
Tweets with profile language Indonesian*	3,968,271 (29.02% of all tweets)	640,466** (19.33% of all users)
Tweets with profile language English*	8,900,104 (65.11% of all tweets)	2,382,708** (71.93% of all users)
Tweets with user-reported country location	574,484 (4.20% of all tweets)	62,698 (1.89% of all users)
Tweets with Indonesia as user-reported country location*	401,185 (2.93% of all tweets, 69.83% of tweets with user-reported country)	51,264 (1.55% of all users, 81.76% of users among tweets with user-reported country location)
Geo-tagged tweets (UberTwitter)	97,308 (0.71% of all tweets)	21,746 (0.66% of all users)
Geo-tagged tweets (UberTwitter) with coordinates in Indonesia*	90,768 (0.66% of all tweets, 93.28% of geo-tagged tweets)	20,089 (0.61% of all users, 92.38% of users among geo-tagged tweets)

*Being geo-tagged from the UberTwitter app with coordinates in Indonesia and identifying Indonesia as the user-reported country location are separate tweet metadata parameters (and a geo-tag from Twitter is separate from both those parameters, although as mentioned above our set had no such tweets). However, only 3,590 tweets have both fields filled. Of these, 3,217 of these have coordinates in Indonesia and list Indonesia as the user-reported country location, and 78 have coordinates outside of Indonesia and a country other than Indonesia as the user-reported country location. The rest are discrepant, with 85 having coordinates in Indonesia but a country other than Indonesia as the user-reported country location, and 210 having Indonesia as the user-reported country location but with coordinates outside of Indonesia. Given the small number of such tweets, we count them along with the set of tweets reported as from Indonesia.

For disaster management, this suggests that a data collection that pulls data using both a bounding box and terms will generate data whose center is in the location of concern and in the language of that population; however, the data will not be exclusively from that region. Thus, this collection strategy will require substantial data cleaning. The second implication is that there is likely to be a significant Twitter population that is bi- or possibly multi-lingual (as is evidenced by both the multi-lingual tweets and the fact that the profile and tweet language are different). Such individuals may be mobilized to help get information to multiple linguistic communities.

We observe that there are only 3,590 tweets that are both geo-tagged from UberTwitter and have a user-reported country location. This suggests that geo-tagged coordinates and the user-reported country location are nearly independent ways of identifying tweets as being in the region of interest. Unless the geo-tagged tweets are spoofed or somehow falsely tagged, they are guaranteed to be in Indonesia so they are more reliable (it may be easier to spoof via a third-party application such as UberTwitter, so geo-tagged reported through Twitter would be most reliable, but we make an assumption that deliberately spoofing a specific geo-tagged coordinate is unlikely—at least unless having a geo-tag becomes or has become a marker of credibility that spammers would seek to exploit); however, there is a far larger volume of tweets with Indonesia as the user-reported country location, presenting a trade-off between volume and accuracy from these two ways of identifying tweets from Indonesia. For disaster management, this suggests that there is value in doing two data collections; one focused exclusively on the bounding box and one focused on the terms and the bounding box so as to get access to the information from those in country but without the coordinates of the geo-tag turned on. In this case, though, care will be needed to de-duplicate the data draws. The decahose is less useful, particularly for less major cities, as there simply will not be sufficient data.

4.1.2 – Users (The Users)

During this time period for Indonesia, the standard deviation in number of tweets per user is huge suggesting an immense skew in usage patterns. The actual skewed distribution of tweets per user is shown in Figure 17, where we see that the majority of users hardly tweet, and a small number of users have a large number of tweets. This is typical of most Twitter data sets. In this data set, 1,660,575 users only have one tweet; whereas, the top user had 33,760 tweets. The top four users[15] in our data from this time period have since had their accounts suspended by Twitter, which suggests that the many have been spam or bot accounts which were discovered. Certainly, high levels of activity is one marker of a potential bot. The top user from our data whose account is still active is @Aquafloater, with 9,080 tweets over the week. This account is a commercial one, tweeting out only links with hashtags of a magazine.

In Figure 1, we see that whether we examine all users, just verified users, or other categories we get approximately the same type of distribution. Note, tweet in Bahasa (the Indonesian language) account for 62.40% of all tweets, and have a patter similar to that of all tweets. There is a peculiar jump in all these distribution plots around 15 tweets. This is most likely a function of the Twitter API, rather than a regularity in Indonesian tweeting behavior. Based on the distribution alone, there is no natural discontinuity from which we might build a usage typology that is more meaningful than frequent or infrequent user.

[15] @BANDUNG_24JAM (33,760), @AllFoodAlltheTi (18,557), @BOTartri2 (10,754) and @BOTartri1 (10,502).

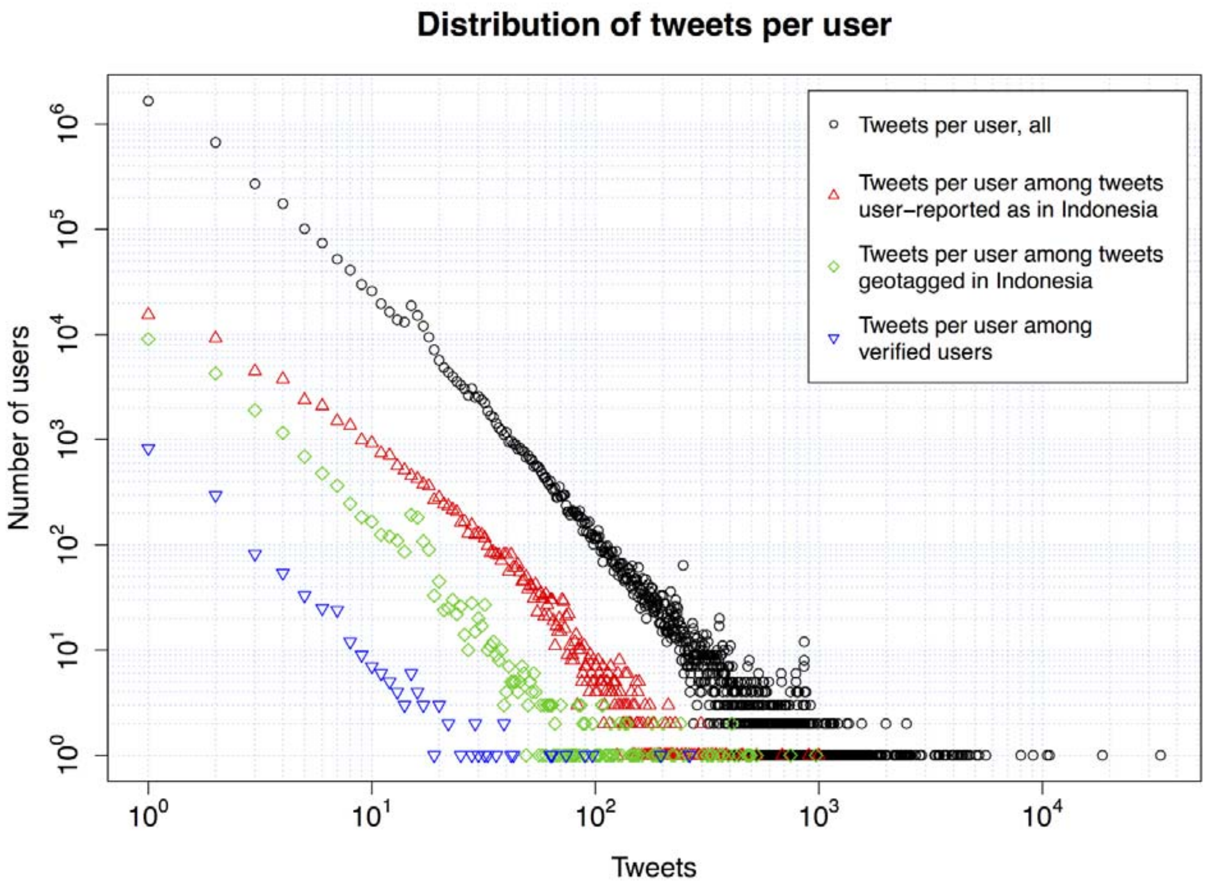

Figure 17: Distribution of tweets by type of user.

For disaster management this distribution means that the identification of influential users or opinion leaders should take into account tweet frequency, in part as low frequency users may not be on-line at the critical time. It also means that, there will be a few users who will have disproportional impact in spreading messages.

4.1.3 - Language

As a note about Twitter data, there are two language metadata tags associated with each tweet: the language of the tweet (i.e. the language to which the user's operating system is currently set when the tweet is written, a setting which affects things like spell check and, if applicable, keyboard mapping, character set, and writing direction), and the language of the user's profile, which the user chooses and which determines what words the user sees in the Twitter interface (a user whose profile is set to Bahasa Indonesia (aka Bahasa Malay) will see at the top of Twitter the words "Beranda," "Notifikasi," "Temukan," and "Saya," while a user whose profile is set to English will see "Home," "Notifications," "Discover," and "Me").

We now consider differences in tweets by language. See Tables 5 (Tweet analysis) and 6 (user analysis). Most of the tweets in this data set are in Indonesian, and most of the users in the data set are tweeting in Indonesian. In some ways, this is not surprising given that these tweets are dominantly from Indonesia and were often selected via terms many of which were in Indonesia or were Indonesian place terms or celebrities. English is the second dominant language. After Indonesian and English, there is Tagalog, the language of the Philippines, and which is consistently the third most represented language (which we include to show how other languages have vanishingly small representation). Indonesian and English are the dominant languages by users also - see Table 6. There are several reasons for the prevalence of Tagalog. First, there are about 5000 Filipinos living in Indonesia, most of whom live in Jakarta, are well educated and who hold higher paying jobs. Thus, they may represent a disproportionate number of the tweeters. Second, the bounding box for Indonesia, overlaps polygons associated with Malaysia and the Philippines. So tweets from these regions may be present. Third, many of the tweets captured by the key terms in our selection list may represent issues of relevance to all the countries in this broader Pacific region. Note there are a few tweets where the tweet language is Malay – but this is more a matter of civic pride than representing a truly different language from Indonesian.

Table 5. Tweet volume by tweet language.

Tweet Language	Tweets (% of all tweets)	Tweets with user-reported country location (% of all tweets with user-reported country location)	Tweets with user-reported country location in Indonesia (% of all tweets with user-reported country location in Indonesia)	Geotagged tweets (% of all geotagged tweets)	Geotagged with coordinates in Indonesia (% of all geotagged tweets with coordinates in Indonesia)
Indonesian	8,530,352 (62.40%)	376,787 (65.59%)	291,641 (72.69%)	81,205 (83.45%)	76,787 (84.60%)
English	4,272,018 (31.25%)	73,832 (12.85%)	32,167 (8.02%)	11,860 (12.19%)	10,189 (11.23%)
Tagalog	158,797 (1.16%)	31,993 (5.57%)	16,008 (3.99%)	1,136 (1.17%)	1,049 (1.16%)

A consistent and interesting pattern can be seen in Tables 5 and 6: the percentage of users who tweet in Indonesian is smaller than the percentage of tweets in Indonesian. This means that users who tweet in Indonesian do so much more frequently in our collected data than do users who tweet in other languages. We also find that the percentage of users who tweet in Indonesian is much higher in the geotagged tweets with coordinates in Indonesia than in other data, pointing to a systematic difference within this set in language behaviors. While this is not surprising, it is surprising that the users who self-report Indonesia as their country location are far less likely to send out tweets in Indonesian than are users who have geotagged tweets with coordinates in Indonesia. Possible explanations could be that those who self-report Indonesia as their country are more educated and more fluent in English, or in jobs that require English usage, or are non-Indonesians living in Indonesia. Another possibility is that such individuals are international travelers or workers, who have at least temporarily, changed their profile to indicate their current location.

We can see that for the subsets that are identified as being in Indonesia either by the user or by the geotag, there is a much higher proportion of tweets in Indonesian. As noted earlier, we should not assume that tweets from Indonesia are likely to be in Indonesian, as English is considered a prestige language within Indonesia (Lauder, 2008). The utilization of English for tweeting may be exacerbated by the fact that most Indonesian tweeters are young, have been in a school system requiring them to learn English, and/or are highly educated and in jobs requiring English. That being said, in subsets, users who tweet in Indonesian are the majority, e.g., in geotagged tweets with coordinates in Indonesia, users who tweet in Indonesian only are the majority at 71.23%. This increases our confidence that most of these Indonesian language tweets are actually from Indonesia.

Table 6. User volume by tweet language

Tweet Language	Users (% of all users)	Users within tweets with user-reported country location (% of all users within tweets with user-reported country location)	Users within tweets with user-reported country location in Indonesia (% of all users within tweets with user-reported country location in Indonesia)	Users within geotagged tweets (% of all users within geotagged tweets)	Users within geotagged tweets with coordinates in Indonesia (% of all users within geotagged tweets with coordinates in Indonesia)
Indonesian	1,515,698 (40.23%)	50,663 (37.15%)	43,818 (41.99%)	18,176 (69.64%)	17,259 (71.64%)
English	1,732,130 (45.97%)	19,826 (14.54%)	12,722 (12.19%)	5,098 (19.53%)	4,291 (17.81%)
Tagalog	98,570 (2.62%)	12,420 (9.11%)	8,822 (8.45%)	652 (2.50%)	591 (2.45%)

As noted earlier, many users tweet in multiple languages. We can use these users and their tweets to look at 'bilinguality' (or multilinguality). This analysis is shown in Table 7. In the total collection of tweets, out of the 3,312,550 users, 330,040 (9.96%) have tweets in more than one language. Of these, 74,661 have tweets in three or more languages, all the way up to an account that sent out tweets in 40 languages. This latter account has been since suspended, suggesting that it was flagged as a bot.

In the full data set, there are more users who tweet monolingually in English than there are users who tweet monolingually in Indonesian. Surprisingly, considering the number of users who tweet in Indonesian but have their profile language set to English, there are almost no users who tweet bilingually in Indonesian and English (at least not in terms of switching their system language—they may well include both languages in a single tweet without switching their system settings).

Different subsets of the data show different patterns suggestion different Twitter behavior for different socio-linguistic groups. The number of English-only tweeters is far lower than in the full set of tweets, although higher in the set of geotagged tweets than in the set of tweets with a user-reported country location. The number of bilingual Indonesian/English tweeters also rises through the subsets, but still being less than 10% of the total population. Interestingly, among the users in the set of tweets with a user-reported country location, both Indonesia as that location or another country, have a very high proportion of users who send out tweets in Indonesian and one or more other non-English languages. The majority of these cases have either Tagalog as the other or another language, and after that, Undefined as the other or another language. Thus, it appears that users in these two sets have a predilection to send out tweets in Indonesian and Tagalog (and possibly others) more than to send out tweets in Indonesian and English. However, what this means is puzzling, as the Filipino population of Indonesia is very small,[16] nor is there an exceptional amount of commerce between Indonesia and the Philippines such that there might be many cross-border companies tweeting in both languages. This is a point that needs further investigation, especially to understand what it would mean to use the set of tweets with Indonesia as the user-reported country location as a means of identifying tweets from Indonesia.

Table 7. Indonesian/English bilinguality and multilinguality in tweets across users

Tweet language	Users (% of all users)	Users within tweets with user-reported country location (% of all users within tweets with user-reported country location)	Users within tweets with user-reported country location in Indonesia (% of all users within tweets with user-reported country location in Indonesia)	Users within geotagged tweets (% of all users within geotagged tweets)	Users within geotagged tweets with coordinates in Indonesia (% of all users within geotagged tweets with coordinates in Indonesia)
Indonesian only	1,247,898 (37.67%)	23,716 (37.83%)	21,910 (42.74%)	14,992 (68.93%)	14,307 (71.23%)
English only	1,502,056 (45.34%)	4,690 (7.48%)	2,188 (4.27%)	2,628 (12.08)	2,026 (10.09%)
Indonesian +	120,998	3,712	2,988	1,736	1,611

16 http://web.archive.org/web/20100818175825/http://www.ops.gov.ph/indonesia2001/backgrounder.htm last accessed – Feb. 2015.

English	(3.65%)	(5.92%)	(5.83%)	(7.98%)	(8.02%)
Indonesian + English (+ other)	36,204 (1.09%)	2,952 (4.71%)	2,306 (4.50%)	425 (1.95%)	390 (1.94%)
Indonesian + English (+ others)	16,541 (0.50%)	7,000 (11.16%)	4,645 (9.06%)	147 (0.68%)	136 (0.68%)
Indonesian + other(s)	94,056 (2.84%)	13,282 (21.18%)*	11,968 (23.35%)†	875 (4.02%)	815 (4.06%)
English + other(s)	56,330 (1.70%)	1,472 (2.35%)	595 (1.16%)	161 (0.74%)	127 (0.63%)

* The most common languages among these others is Tagalog (present in 5,002, or 37.66%, of these users) and Undefined (4,522, or 34.046%, or these users).
† The most common languages among these others is Tagalog (present in 4,283, or 35.79%, of these users) and Undefined (4,236, or 35.39%, or these users).

This analysis is useful in providing a baseline for disaster management. It demonstrates that when using a bounding box and terms, the tweets and users identified may be from regions other than the one of interest, and the population of interest. Within this broader signal, however, there will be a smaller set of users and tweets of relevance. Extracting these will take substantial post-processing focusing on the language and the geographic information. This problem will be reduced for those tweets gathered by a bounding box only, but not eliminated due the Twitter API pulling data based on overlapping polygons. It also demonstrates that although Indonesian is the dominant language, and dominates Twitter in Indonesia, that messages in Twitter are reaching and coming from those who speak English and Tagalog. Thus, authoritative messages should be retweeted in all three languages for more comprehensive coverage. A third implication of this analysis is that there is a small but consistent multi-lingual population. There is a possibility of mobilizing this group to spread early warnings. A fourth implication is that due to the high fraction of those in Indonesia tweeting in Indonesian, Twitter may provide better coverage of this population than other linguistic minorities.

4.2 Spatio-temporal Patterns

We now ask, what are the spatio-temporal patterns for tweeting? For disaster management, even if a media is used, if it is not being used at the time or in the region where a disaster hits, it may be less valuable. Thus, this analysis suggests the spatio-temporal weaknesses and strengths in the Twitter coverage. Knowing when the population normally tweets and from where is useful in determining the severity of impact after a disaster event; i.e., if a different temporal pattern ensues or if formally high tweeting regions are conspicuously silent, that is a signal of impact.

4.2.1 - Dynamics

To orient the discussion of dynamics we first examine the broad trends based on the decahose data for all of Indonesia. In this case, using only those tweets whose geo-tag places them within Indonesia, we see a general pattern of tweeting with strong temporal regularities. These values are based on data from Jan 1 2011 to August, 2014. The tweets have a characteristic daily (Figure 18) and weekly (Figure 19) pattern. In Figure 18, the mean number of tweets by time of day is shown with "00" being midnight in Indonesia. On average, most tweeting is done between 6 pm and 10 pm. However, there is a significant amount of tweeting during the work/school day. While, there is significant variance, the maximum tweets in an hour follow the same basic pattern. Although midnight is general a lower tweeting hour, the maximum number of tweets seen in any hour did occur at midnight. The number of hashtags used per hour of day follows a similar pattern – so as Indonesians tweet more they are also tweeting about a greater variety of things.

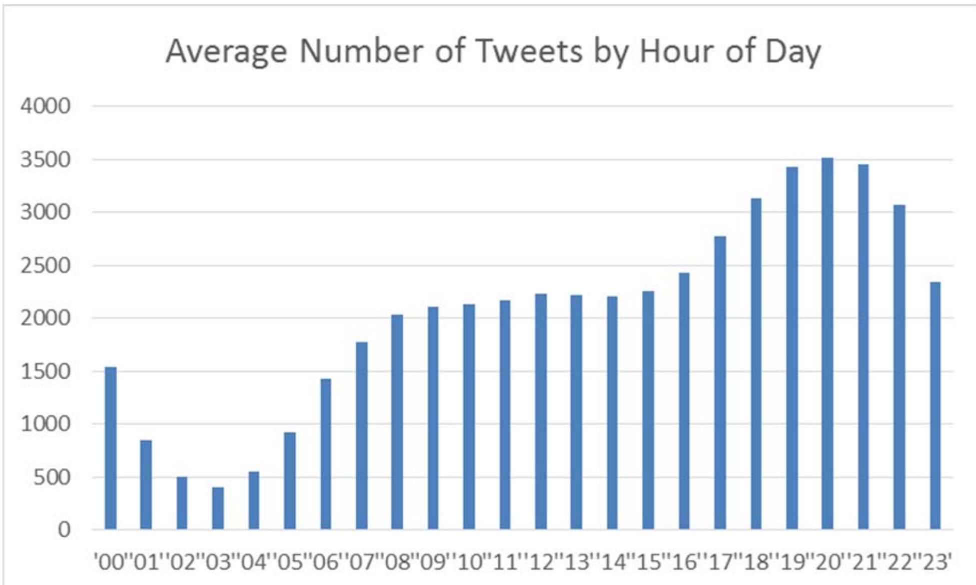

Figure 18: Average number of tweets per hour of day, for only geo-tagged tweets from all of Indonesia (00 is midnight in Indonesia) – decahose

Using the decahose data, we do see that in general, most tweets are occurring on Sunday – Figure 19. Interestingly, in 2011, Thursday was the busiest tweeting day for Indonesia.[17] This was followed by Monday, then Sunday and Saturday. This may reflect changes in access to the internet. Although the warnet, the internet cafes, were used early on, much of the original access was in universities and corporations. So a Monday-Thursday pattern might have reflected tweets from work and school. An alternative explanation is that with the growth in number of users and the increased use of cell phones to tweet, Indonesians may be increasingly using Twitter to socially coordinate and gossip about their weekend entertainment. Neither explanation can be ruled out. Consistent with the 2011 data, we find the least activity on Tuesday and Wednesday on average. However, the maximum value per day was observed on a Wednesday. The distribution of hashtags by day of week follows a similar pattern, with Sunday being the day on average when the greatest variety of things are discussed.

[17] http://www.thomascrampton.com/Twitter/indonesia-Twitter-stats/ last accessed – Feb. 2015.

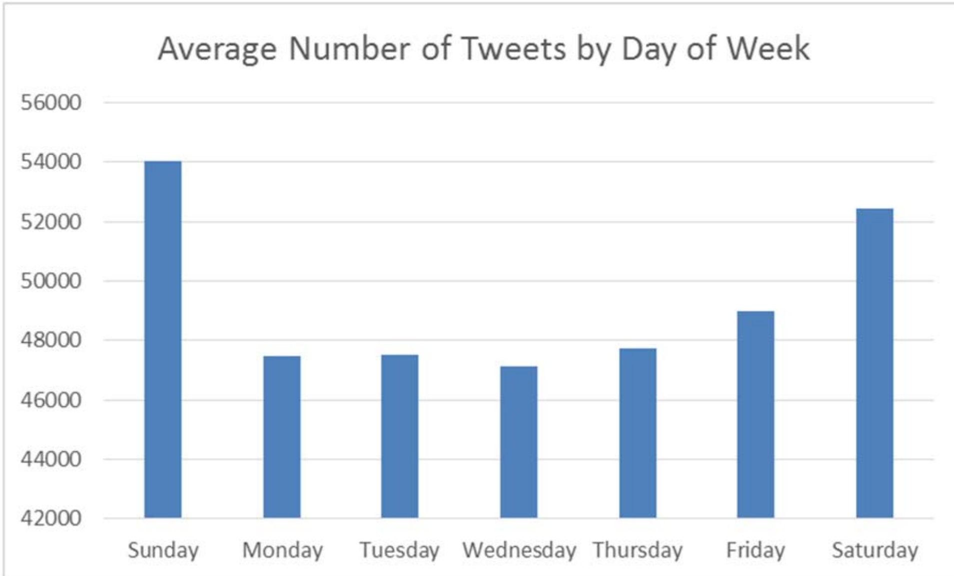

Figure 19: Average number of tweets per day of week, for only geo-tagged tweets from all of Indonesia – decahose

There is significant variation, so in any given week, Sunday may not be the highest. Moreover, for users discussing disaster related information a different pattern may be in effect. Further, we suspect that this decahose data is largely driven by Jakarta. Our analysis reveals a different pattern in Padang with Tuesday being the highest tweeting day in CS4 and Wednesday in CS5. Also in Padang, the work/school hours are generally higher than the evenings.

A more in-depth look using the Indonesia tweets collected CS3 clarifies this point. In Figure 20 the number of tweets per hour for a typical week in 2013 is shown. In this example, the most tweets are occurring on Saturday. The second pattern to note is that while the total collection of tweets cannot be reliably said to be from Indonesia, we see the same pattern in the tweets that are specifically geo-tagged within Indonesia. Similar patterns exist for the number of users by hour, and the number of geo-tagged users. The weekend spike is not as pronounced for users as it is for tweets, which indicates that daily and hourly differences are more a function of specific users being more active rather than more users being active. The implication for disaster management is that a data collection scheme such as CS5 that focuses on geo-tagged tweets is reasonably representative of the general dynamics of the entire population.

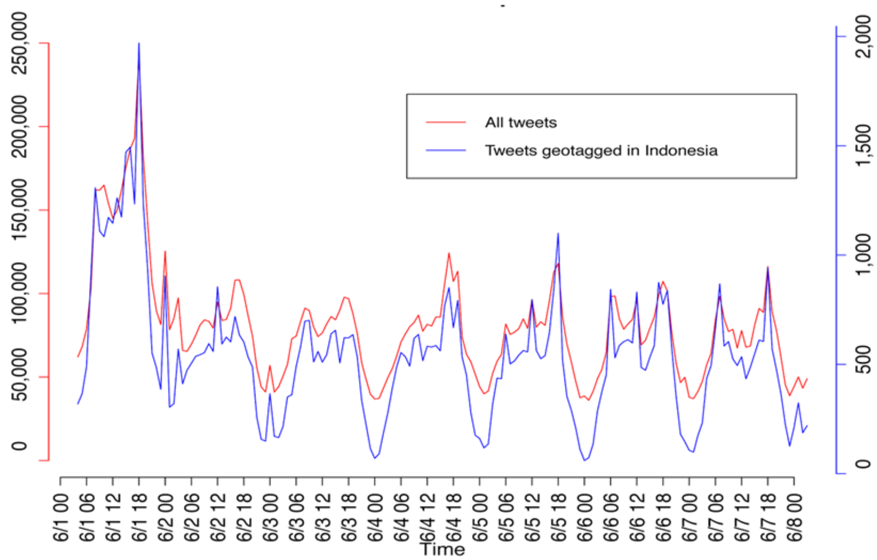

Figure 20: Temporal signature of Twitter for June 1-June7 2014 showing tweets per hour for all tweets and geo-tagged tweets.

When looking at the Tweets in CS3 by hour for the different language patterns (Figure 21), we find that the English language tweets do not follow the same pattern as the Indonesian language tweets. This suggests that some of these tweets are coming from users with different temporal patterns of activity, most likely because they are in a different time zone. The geo-tagged English language tweets with coordinates in Indonesia have a temporal signature that is similar to that of the Indonesian language tweets. The number of geo-tagged English language tweets with coordinates outside of Indonesia is too small to provide a meaningful comparison. The implication for disaster management is that non-geotagged tweets with a different temporal signature are likely to not be from the region of interest.

There are several implications of this temporal analysis for disaster management. First, Twitter will be less valuable for reaching the population with an early warning on weekdays and between 11 pm and 6 am. Second, the rhythm of Twitter usage appears to be related to the overall Indonesian culture and not linguistically based. Hence, messages do not need to be staged by linguistic group. Third, even though there is a temporal pattern, there appears to always being someone on-line. Efforts to ensure these individuals attend to messages from the authoritative sources and then communicates to the population via non-Twitter means may be an effective warning strategy. Fourth, the distinctiveness of the English tweets suggests that they many of them may not be from Indonesia. Thus knowledge systems for disaster management might want to use the temporal pattern of Twitter as a way of ruling out tweets that are not relevant.

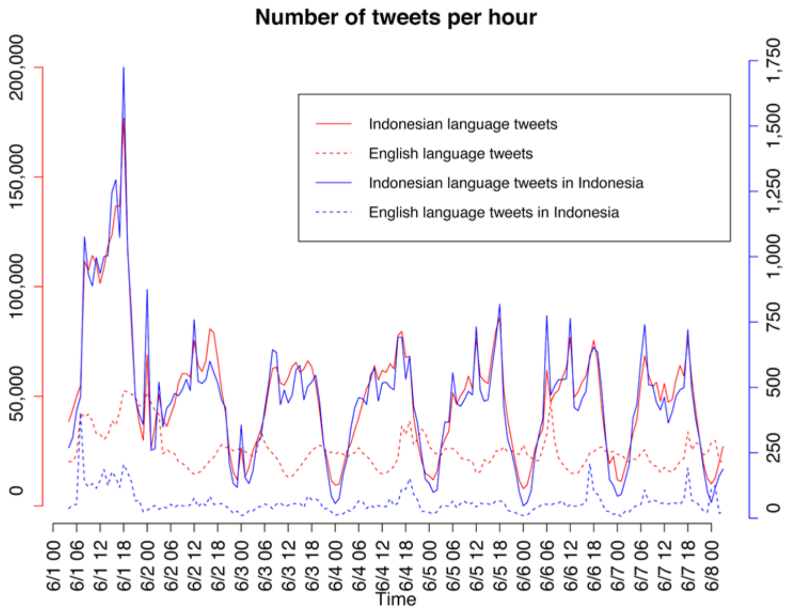

Figure 21: Number of tweets per hour for June 1-June7 2014 showing language differences and similarities.

4.2.2 - Geographic distribution

In CS3 there were tweets from users giving their location as in one of 69 different countries. And there were tweets with geo-tags that were far outside of Indonesia. User-reported locations are shown in Table 8 based on CS3.1 data. There are more tweets and users from nearby countries. In part this is due to the nearby countries having bounding-boxes that border or overlap with that for Indonesia. Of these countries, the Philippines stands out as it is the furthest from Indonesia, and has the least overlap. This suggests that tweets by users listing their location as Philippines may be from Filippinos currently and perhaps temporarily working in Indonesia.

Table 8. User-reported country location

Country	**Tweets**	**Users***
Indonesia	401,185	51,264
Malaysia	154,035	9,001
Philippines	7,420	303
Singapore	5,545	442
Thailand	3,144	148
United States	1,218	826
Australia	571	50
Great Britain	427	273
France	135	96
Total	574,484 (4.20% of all tweets)	62,698 (1.89% of all users)

* There were 222 users with tweets with two different countries (i.e., they changed their country designation during the collection period), 6 users with three countries, and 2 with four countries. There were 120 such multiple countings for users with tweets in Indonesia, 129 for Malaysia, 21 for the United States, 92 for Singapore, 3 for the Philippines, 15 for Great Britain, 26 for Thailand, 6 for France, and 5 for Australia. The 62,698 number is the correct total and does not include the multiple countings of the same item.

For geo-tagged tweets, we can plot their coordinates over a map of Indonesia or as a heat map showing densities using our Twitter disaster tool (Landwehr et al, this issue). An example for the broad data set is shown in Figure 22. During our collection period we observe tweets in all regions colored blue or green; however, there are more tweets in the green areas. Unsurprisingly given its population density, many of the tweets are concentrated in Java. Most tweets are from high population centers; specifically the large cities such as Jakarta (the 2011 Twitter capital of the world). This finding suggests that Twitter may be useful in these population centers for providing general situation awareness, providing early alerts, assessing damage (location and severity), and coordinating response and education. It may be less useful in extremely low population centers for identifying where people are due to the paucity of tweets. However, given the increased use of social media in Indonesia in general, this situation is likely to change in the near future.

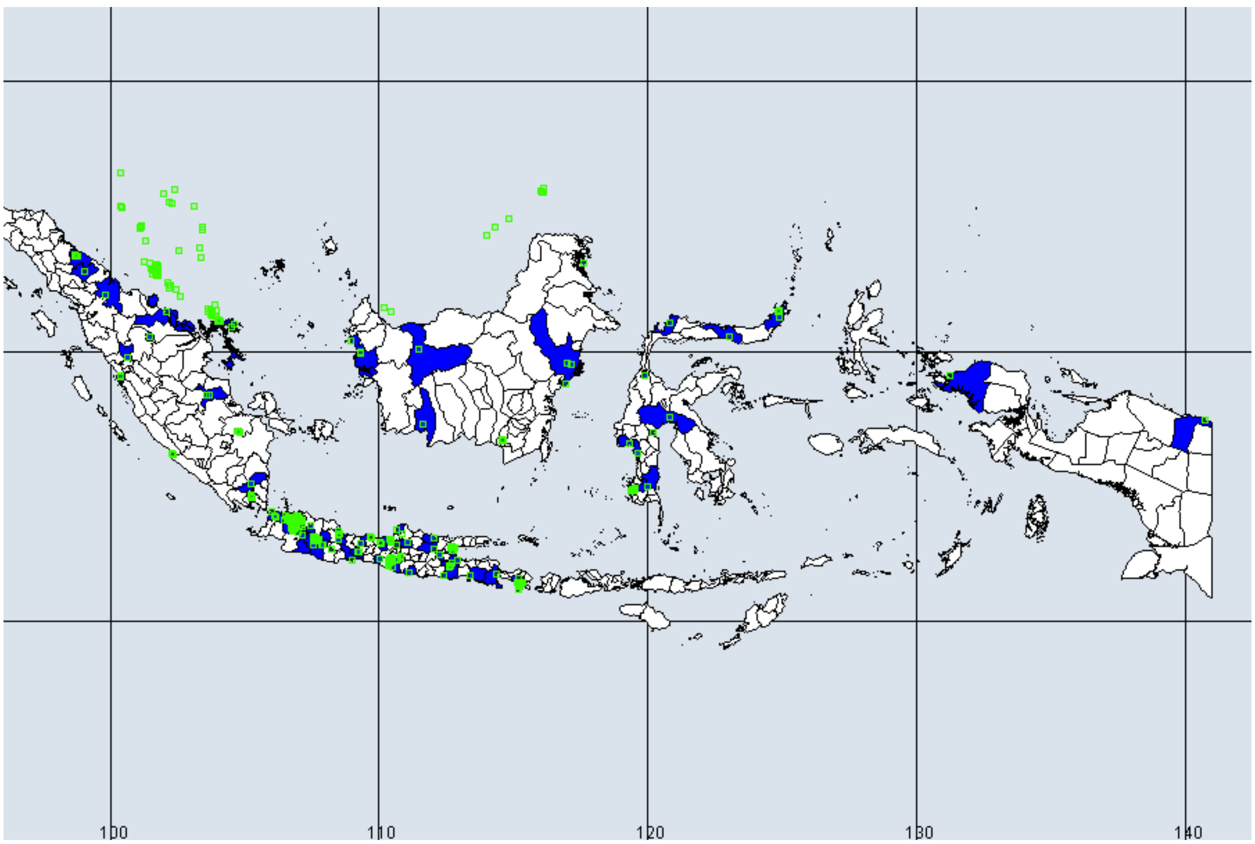

Figure 22: Geographic mapping of geo-tagged tweets in Indonesia.

The tweets are themselves not uniformly distributed within cities. For example, in Figure 23, which contains just the geo-tagged tweets for Padang, we see high levels of tweeting along the coastal areas and less internally, as the countryside becomes harder to traverse or less developed. In Figure 24, a heatmap is shown, showing the fraction of tweets within the political sub-regions in Padang (CS5 using the data cumulative through the beginning of August 2014). This image shows that the majority of the tweets are coming from three areas, which are the locations of the universities. We also find strong spatial variance in where tweets are originating from by time of day and day of week.

There are several implications for disaster management. First, the decahose is insufficient for assessing population locations. Second, collections focused on only a bounding box provide better coverage of the region of interest; however, the downside of using only geo-tagged tweets is that it misses most of the user population. This is important for disaster management as those missed tweets may contain critical information about who needs what kind of help. To overcome this limitation, work is being done on inferring location based on the ties among users (Compton et al., 2014), the time of day implying certain time zones (Mahmud et al., 2014), or even the language. Doing such inference is still in its infancy. Third, the Twitter population while more likely to be in cities, may not be in the most populated areas of the city. Fourth, although systematic data on age is not available with the tweets, assessing the spatial pattern may provide insight into which segments of the population can be reached by this media.

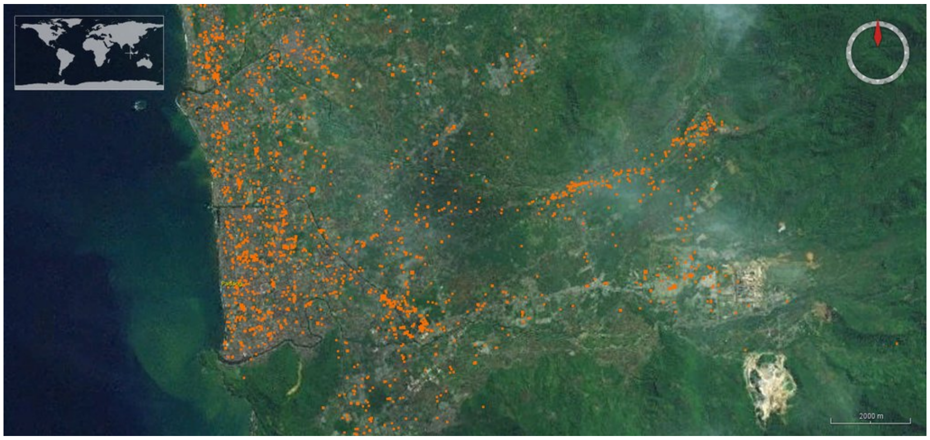

Figure 23: Geographic mapping of geo-tagged tweets for a random Monday in Padang in 2014.

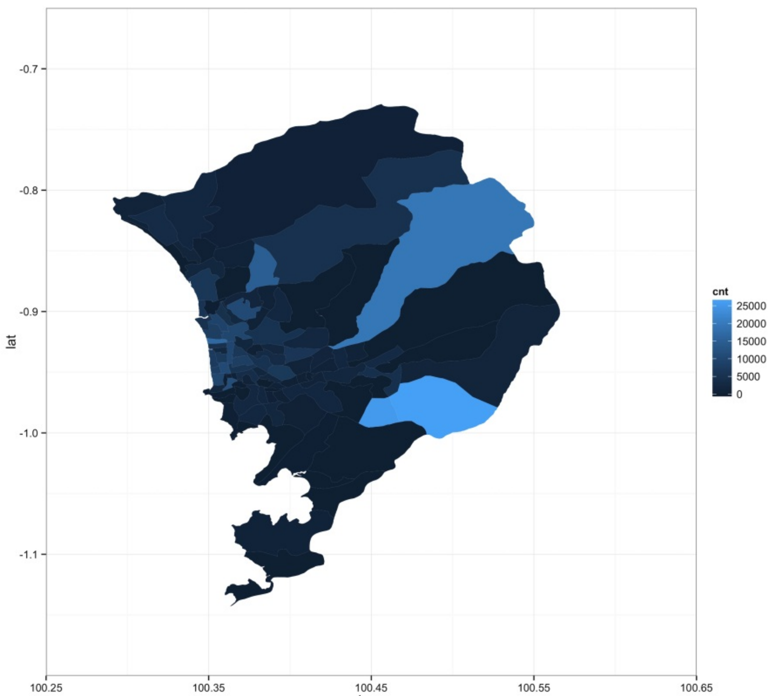

Figure 24: Heatmap of geo-tagged tweets for Padang, using CS5.

The most detailed spatio-temporal patterns require focused local collection like CS5. This data can support evacuation route planning by providing insight into where people are when. However, given the relatively small fraction of tweets that are geo-tagged it is likely not to be sufficient for identifying all those who need to be evacuated. Whether the temporal trails used by those who are geo-tagging there tweets provide guidance for evacuation is a subject for future study and beyond the scope of this paper.

4.3 – Identification of Opinion Leaders

There are multiple networks inherent in Twitter data. These include, but are not limited to: the *mention* network, which is a directed network of one user mentioning another in a tweet using the "@" symbol; and the *hashtag* networks, which are undirected networks where users are connected if they both use the same hashtag. A *hashtag co-occurrence* network can be constructed from the two-mode tweet-to-hashtag network and then projecting that into a 1-mode network of hashtags. Because of the nature of our sampling, which was word-based rather than network-based (e.g., network snowball sampling is requesting all followers of a given user, then request all the followers of each of those users, etc.), we will not go deep into network analysis, instead only considering what would be the most immediate network data findings. Namely, who are the users most mentioned (highest degree centrality), which hashtags appear the most frequently in the tweets, and which hashtags are used by the most number of users. While the hashtag network is derivable regardless of the collection strategy, the mention and retweet network will be overly sparse unless a topical or regional approach is taken. Thus selections from the streaming API are better for identifying opinion leaders than is the decahose.

Two users may not represent independent observations, especially for network behavior (coordination, influence). This overlap in messages will be extremely important in disaster situations in the case of herd behavior that clusters around inaccuracies (such as the case of vigilante online attempts to identify the perpetrator of the Boston marathon bombings), but influence or coordination that exists beforehand will not necessarily hold in disaster situations. Thus the value of investigating potential network mechanisms using tweets, especially given the difficulties in the modeling required for such investigation (Shalizi & Thomas, 2011), is unclear. The data, however, is ideally suited for assessing the structure of extant tweeting networks.

In Table 9 based on CS3.1 we see that among Indonesian language tweets, the list of most mentioned users is similar to the set of all users, probably because Indonesian language tweets make up such a large volume of the total. However, it is noteworthy that the activity from English language tweets was far less than from Indonesian language tweets for the top hashtags to be nearly the same.

Looking at some of the handles: @chocoliciousmks is the handle of Chocolicious, a "premium cookies and cake" bakery in Makassar, the capital of the province of South Sulawesi.[18] @detikcom is the handle of Detik, an Indonesian news site.[19] @radityadika is Raditya Dika, an Indonesian writer, producer, director, and comedian. @alysyarief and aliando26 are handles of Aliando Syarief, an actor in Indonesian soap operas.[20] The Geo-tagged network is dominated by highly central news and political commentary users. In contrast, those who report Indonesia as a country frequently mention stars and various teens and kids, who are prolific users. It may be that the difference between the geo-tagged users and those who report

[18] http://www.chocoliciousmks.com/ last accessed – Feb. 2015.
[19] http://www.detik.com last accessed – Feb. 2015.
[20] http://id.wikipedia.org/wiki/Aliando_Syarief last accessed – Feb. 2015.

Indonesia as a country is age, with the former being dominated by older users and the latter dominated by very young users, i.e., junior high and high school students. We see a similar difference in the hashtags used with the geo-tagged network focusing on hashtags referring to political or economic issues or locations and the Indonesia as country network using hashtags that are more transient and often referring to celebrity events such as birthdays or music releases.

Table 9. Most mentioned users (by tweets)

	Geo-tagged with coordinates in Indonesia		**All**		**Indonesian language tweets**		**User-reported country location of Indonesia**	
Rank	**Handle**	**#**	**Handle**	**#**	**Handle**	**#**	**Handle**	**#**
1	pilaradio886fm – news/music	869	Chocoliciousmks - bakery	89,129	Chocoliciousmks - bakery	89,127	Alysyarief – actor Aliando Syarief	843
2	Akuekaaj - Selamatkan Islam	819	Detikcom - news	75,352	Detikcom - news	72,988	aliando26 – actor Aliando Syarief	516
3	NasDem – political party	278	Erlanhandsomnia - Erlan Pratama, Bogor	30,326	Erlanhandsomnia - Erlan Pratama, Bogor	30,325	PrillyBie - Prilly Latuconsina teen idol	507
4	Moharifwidarto – commentator		Pendingmoveon	30,309	Pendingmoveon	30,308	Iqbaale – child idol	489
5	COVtranstv – Celebrity On Vaction Org	220	Radityadika – Indonesian writer	25,149	Radityadika – Indonesian writer	24,474	AlvaroMaldini1- child idol	293
6	fathayu - Safatah Purwonoto	213	Gerindra – Great Indonesian Movement Party	24,940	Gerindra – Great Indonesian Movement Party	23,715	jokowi_do2 - Joko Widodo - president	254
7	Acenaradas - commentator	187	Afifaahk - Afifah Kamilia	21,254	jipanks_ - Irfan Gunawan - fashion	21,254	GGSerigalaSCTV_ - news	240
8	99ersRadio_Bdg – news/radio	179	jipanks_ - Irfan Gunawan - fashion	21,254	Afifaahk - Afifah Kamilia	21,254	Prabowo08	238
9	Aries_tresna - commentator	170	Prabowo08 - blogger	20,603	Prabowo08 - blogger	19,792	Cassandrasleee – teen idol	196
10	AngeliaCristin_ - commentator	162	Gabriele_Corno – business man	20,349	TrioMacan2000 – controversial tweet site	12,730	Salshaabilaa - Salshabilla Adriani – teen idol	196

Looking at hashtags by tweets (Table 10) and hashtags by users (Table 11) does not reveal strikingly different patterns, but the number of times the top hashtags appear in all tweets versus among all users is an order of magnitude different. Specifically, the 735,354 tweets with the hashtag #gameinsight were sent out by only 24,905 users, suggesting a concentrated campaign (possibly including spam). The gaming and mobile device/tablet theme of the other tweets seems to also point to some sort of promotional campaign and/or to spam.

In contrast to who is mentioned (Table 9), the most mentioned hashtags by tweets in CS3.1 (Table 10) or users in CS3.1 (Table 11) is not very similar between the All and the In Indonesian Language. This may be because those users tweeting in Indonesian are dominating the mentions network (they mention the most), but not the hashtags (more users regardless of sub-group use hashtags).

Table 10. Most used hashtags (by tweets)

	Geo-tagged with coordinates in Indonesia		**All**		**In Indonesian language tweets**		**User-reported country location of Indonesia**	
Rank	**Hashtag**	**#**	**Hashtag**	**#**	**Hashtag**	**#**	**Hashtag**	**#**
1	NP	312	gameinsight	735,354	WeAreARTRILOVERS	21,254	NP	1,297
2	jakarta	161	androidgames	419,389	News	16,517	Happy8thAnniversaryELF	557
3	Indonesia	87	android	416,438	GolkarforPrabowoHatta	16,433	Np	204
4	IndonesiaBangkit	83	ipadgames	235,081	ARBdukungPrabowoHatta	10,810	OK	173
5	LMen	65	ipad	233,134	Tuit	10,530	WeWantBDLBack	171
6	YUPITER	65	food	89,420	Jakarta	10,504	SoalnesianKuis4	157
7	ONEFORALL	64	iphonegames	80,666	Video	9,725	AliandoPrillyCepetJadian	139
8	Bali	46	iphone	80,401	Indonesia	9,406	np	129
9	JKWJK	44	jakarta	29,700	jakarta	8,913	ChiBiDay	124
10	bandung	41	Indonesia	23,134	indonesia	8,130	Smile4Us	107

Within other subsets, a hashtag that appears in many different capitalizations is #NP, which can stand for "now playing" or "no problem."[21] Informal exploratory qualitative investigation of tweets shows a practice of users tweeting out #np followed by a song name. #Happy8thAnniversaryELF is a hashtag for the 8th year of operation of E.L.F., the official fan club of the Korean pop group Super Junior.[22] #WeAreARTRILOVERS is something Indonesian (it sometimes appears with #BersamaARTRILOVERS, and bersama may be translated as "we are") but it is unclear what. #GolkarforPrabowoHatta is, unsurprisingly, a political hashtag; Golkar is a politician and Prabowo-Hatta a political party.

In examining Tables 10 and 11, few disaster related terms show up. There are several reasons for this. First, this is data from the CS3 data collection for the months of June and July 2013 when there was no disaster. Second, the set of terms was very broad and included most major city terms as well as disaster terms. This was to try to force more of the tweets to be from

[21] http://tagdef.com/np last accessed – Feb. 2015.
[22] http://www.allkpop.com/article/2014/06/super-junior-celebrate-the-8th-anniversary-of-official-fanclub-elf last accessed – Feb. 2015.

Indonesia. For disaster management, the implication is that use of city names does not guarantee coverage of the interested area. Further, under non-disaster conditions, disaster terms are very infrequent. Hence, specialized collections may be needed just as the disaster begins so as to capture disaster specific tweets. Across all data collections there are very few tweets that contain disaster terms and that are meaningful from a disaster management perspective. Developing a strategy to extract such tweets would be good for disaster management but is beyond the scope of this paper.

Table 11. Most widely used hashtags (by user)

	Geo-tagged with coordinates in Indonesia		**All**		**In Indonesian language tweets**		**User-reported country location of Indonesia**	
Rank	**Hashtag**	**Times used**	**Hashtag**	**Times used**	**Hashtag**	**Times used**	**Hashtag**	**Times used**
1	NP	248	gameinsight	24,905	WeAreARTRILOVERS	19,030	NP	1,178
2	Jakarta	116	food	21,298	GolkarforPrabowoHatta	16,433	Np	175
3	ONEFORALL	64	inspiration	15,499	News	15,428	Happy8thAnniversaryELF	166
4	LMen	56	CGE	15,282	ARBdukungPrabowoHatta	10,810	OK	158
5	Indonesia	55	Indonesia	13,465	Jakarta	9,650	SoalnesianKuis4	142
6	YUPITER	52	androidgames	13,106	Tuit	9,537	np	119
7	jakarta	51	android	12,320	Video	8,634	LM	92
8	indONEsiabangkit	41	Jakarta	11,082	Indonesia	8,446	NowPlaying	84
9	JKWJK	41	sunset	10,519	jakarta	8,078	BISMA	84
10	plsRT	40	floating	10,379	PilihanKuSatu	7,504	Play	83

From a disaster management perspective this suggests that if we wanted to use the Twitter social networks to provide information or do local coordination, then it is likely the key actors will vary by the specific community. We briefly explored this issue with respect to the Padang data. Using just the data for April and May of 2014 from CS5, we constructed a directed network such that a user was linked to another just in case they mentioned or retweeted that other user. There were 5,391 users in the sample and an associated 128,202 tweets with mentions. Of these users only 1,247 actually mentioned or retweeted another. The resulting "network" contains multiple components as can be seen in Figure 25. This suggests that there is no one user who will be automatically serving as a source of information at an informal level (keep in mind that at a formal level many users follow the official BMKG site).

If we focus on the main component, top left in Figure 25, and expanded in Figure 26, we see that the network is not strongly connected. Figure 26 uses the standard Kamada & Kawai (1989) layout algorithm for graphs, but sizes nodes by their degree centrality. By definition, nodes in networks that are in particular elite positions e.g. high in degree centrality or betweenness, are influential and have been shown to act as informal leaders. In principle, if these users can be identified then they can potentially be mobilized. A few actors stand out as being strongly connected to others – high degree centrality in the mentions network (these are the large nodes in Figure 26). These are individuals who are mentioned or mention others (which includes being retweeted and retweeting others) frequently; these are not the actors who tweet the most frequently. The Twitter handles for the top five of these users are: Saibisikuku, siskatejaa, naolarissa, vionarethaa, and FannySelvina. All of these accounts appear to belong to teenage girls. Thus, for at least this sub-community, the informal leaders are teenagers. From a disaster management perspective it may be possible to mobilize these users to support early warning activity; however, since they are teenage girls, they are unlikely to be treated as having an authoritative voice by the local community. This suggests that for disaster management new network metrics for identifying opinion leaders might be needed. We note that tracking followers is unlikely to work, as users frequently "buy" followers or use bots to increase their apparent following.

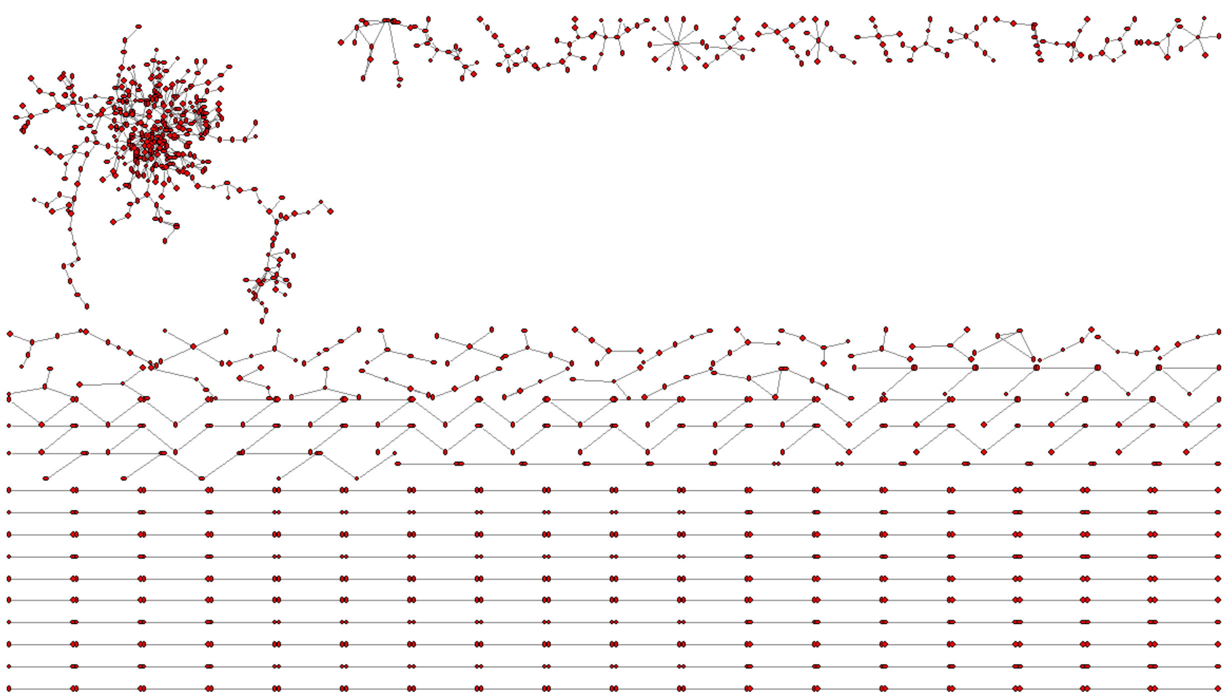

Figure 25: Multi-component network for mentions and retweets in Padang data – April and May 2014.

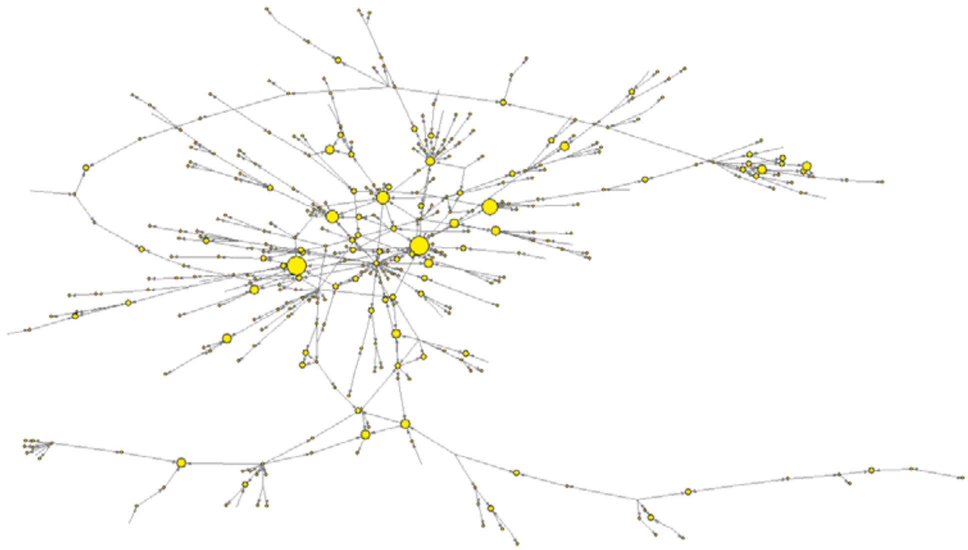

Figure 26: Largest component for Padang with nodes sized by total degree.

5 – Discussion

In a disaster situation, a person using social media as a source of information about the disaster will first and foremost need to know which tweets are actually coming from the region. Other tweets will include a variety of types of information less relevant to disaster management such as calls for donations to relief efforts, calls for help, offerings of prayer, and expressions of solidarity or outrage (Landwehr et al., forthcoming). Such tweets obfuscate who needs what. For location to be useful, either all geo-tagged tweets from the region are needed (which is costly) or location needs to be ascertained by looking at fields that very few users have filled out. Issues of bias and representativeness need to be considered. If the goal is to use these tweets to know where the population is, or to “get the word out” then the lack of representativeness may not be an issue.

This work suggests that, Twitter can be used to reach and do a situation assessment of the major population centers due to better coverage. Its value for remote locations and rural areas is still to be determined. Within cities, the coverage will not be uniform – temporally, geographically, or socio-demographically. Our preliminary work on Padang suggests that not only is this the case, but that the informal leaders of these sub-communities may not be those whom policy makers and first responders are used to working with – e.g., teenage girls. Future work should examine this issue in greater detail and determine if it is possible to automatically find local opinion leaders by group who can help serve as informed correspondents.

In a disaster situation, the volume of tweets may reflect a surge in needs or concern. This work suggests that there is a natural volume that changes by time of day and day of week, against which this surge needs to be calculated. Thus, to use Twitter in a disaster situation may require the responders to have already collected baseline data. This work further shows that in collecting such data a naïve approach to tweet selection cannot be used. Rather, the tweets need to be carefully collected, and then assessed for relevant differences given the linguistic, spatial and temporal subsets. In particular, such a baseline should clearly focus on tweets from the region of interest, but account for the biases inherent in focusing on geo-tagged data. Care must also be taken in inferring location given language.

In these data, the number of tweets with a user-defined country location are more numerous than even the number of geo-tagged tweets reported elsewhere (Morstatter et al., 2013). This, by itself, suggests that in Indonesia, spatial oriented disaster response tools using Twitter may be more effective than in other countries. In these data there is a large fraction of users who tweet in Indonesian and a non-English language or languages, most often Tagalog. The two languages are only distantly related, so this is likely not a case of linguistic confusion or crossover. This may be a sign of the mobility between Pacific Rim countries. In particular, this may be due to Indonesian (or Malay) expats in the Philippines; or, Filippino expat in Indonesia. From a disaster management perspective this has several implications. First, there are natural conduits in Twitter for information to flow informally between the countries which will be valuable for storms that hit multiple countries. Second, disaster management tools in this region that rely on Twitter, must consider at least 3 languages – Indonesian, Tagalog, and English. Third, reaching groups that might not speak the official language, might in some cases be easier due to their presence on Twitter.

As previously noted, social media usages is strongly integrated into daily life in Indonesia. This is reflected in the fact that, as we found, many of the top hashtags are referring to political and entertainments issues, even though we were attempting to select those tweets that were more related to disaster issues. From a disaster management perspective the good news is that there is a community attending to social media and so training is social media usage is not needed; but the bad news is that there may be more irrelevant information to sift through on social media from Indonesia than in many other countries when a disaster hits. Future work should explore these issues and determine the extent to which those Indonesians who are tweeting about and following entertainment celebrities are also following the disaster warning agencies such as BMKG.

In general, Twitter data is filled with commercial activity and spam. In the data we collected, we find relatively low levels of such data (exceptions here are those tweets by the suspended accounts and those regarding chocolicious). This may be due to Indonesian Twitter users not focusing on generating or sending commercial tweets, and instead turning to Twitter as a source of celebrity and entertainment news as often happens in the young adult American Twitter community (Hargittai & Litt, 2011). Alternatively, this may be a bias introduced by the use of disaster related terms for data selection; terms that may be consciously avoided by corporate users. From a disaster management perspective, however, this focus of interest suggests that in Indonesia, co-opting entertainment celebrities to help get the word out would be an effective Twitter disaster response strategy.

The geographic bias towards large population centers is expected, but what is important to note is that there is some activity from provinces other than those in Java. Recalling again that users from which we extracted geo-tags (users who are using UberTwitter) are not any sort of representative sample, but again recalling that in a disaster situation this non-representative subpopulation would be useful in itself as a way to precisely find locations, we see that there is some presence of Twitter users throughout Indonesia. Our data does not contain activity in all provinces. Future work, using either more data streams or gaining access to the entire firehose, is needed to determine the ubiquitousness of Twitter usage in these rural areas. Knowing the extent of not just Twitter usage, but geo-tagged Twitter usage, is important for constructing a disaster management social media platform.

A detailed examination of the users shows diverse communities with their own local opinion leaders. There are high general levels of followership, but a tendency to use Twitter for news and entertainment. We also note a high preponderance of government and disaster service agencies that have Twitter accounts. This suggests that a disaster management system that is based on Twitter, will want to have a very comprehensive engagement strategy. The baseline data for example, should be assessed to identify the kinds of users that are local opinion leaders and to identify the range and geo-region of the different Twitter communities. These local leaders should be engaged as part of an early warning system and response system. This might lead to unexpected new target groups – such as training teenage girls to be information spreaders for the disaster service organizations. Similarly, the government might want to set up a network with entertainment celebrities so that they tweet critical disaster news. Such a system would want to work with the Indonesian disaster services that are already tweeting disaster news such as BMKG. Future work should confirm these strategies and assess the current Twitter networks associated with government and disaster service providers to determine the extent to which these are naturally being followed as a source of news.

6 – Conclusions

An examination of Twitter usage in Indonesia demonstrates the viability of using this data steam for disaster management. We find wide spread usage, high prevalence of geo-information, relative uniformity of the language across the tweets, and active use of and support for Twitter among local officials and official disaster response groups. This suggests that disaster management tools that employ Twitter should be relatively effective in Indonesia. This point is further enforced by the fact that many of the dominant Twitter users are also on other social media platforms, such as Facebook and Instagram, and Twitter is often used in social community settings, which together suggest that disaster information from Twitter will cross into other communication spheres. We find high variance in volume by day of week and hour of day, multi-component mentions networks, and variability by apparent geo-linguistic groups.

This suggests that to be effective a Twitter based disaster management system should first develop a geo-temporal mapping of "normal" Twitter activity. Against such a baseline, it will be easier to assess disaster related activity. Establishing such a baseline will require careful collection and assessment. As most usage is geared toward sports and entertainment, most Tweets will be irrelevant. Even in a data feed selected using disaster related keywords, most Tweets were not concerned with disaster issues and many of the users were potential bots or users misusing Twitter, language was mixed and spatial coordinates were rarely available. Care needs to be taken to maximize the number of relevant tweets collected.

The foregoing analysis suggests the following high level guidelines:

- The socio-cultural-technical context will impact which social media has the greatest utility for disaster management.
- The data collection strategy will impact the value of the data for disaster management. For Twitter, collection for the region and then post-processing for disaster related issues is likely to yield more useful data. Moreover, it generates data with sufficient coverage that one can determine the utility of Twitter in reaching different linguistic, age, and sub-regional groups.
- Knowing the base-line spatio-temporal patterns supports disaster planning, and supports predicting the potential impact of an event.
- Opinion leaders can be identified, however, they might not be viewed as authoritative. The mobilization of these individuals, however, is likely to increase community resilience by providing a mobile force capable of rapidly diffusing information to the

broader community and collecting information on problems and re-broadcasting these to authoritative sources.

In this paper, we have focused on Twitter usage in non-disaster periods. Our objective was to use this data to provide a baseline for Twitter use to support disaster management. The baseline provided herein can be used to compare Twitter usage during a disaster to this non-disaster usage to determine special crisis response modes of communication. This baseline can also be used to establish guidelines on how to collect and use data for disaster management, and to set expectations for what Twitter data can and cannot be used for. While we do not claim that this paper provides an exhaustive catalogue of such use, it does demonstrate that the holes in the coverage of the population, and the limits in using this data as the sole technology for reaching the entire population with information, for determining who is where when, and for identifying those who can be mobilized to support the response effort.

The current disaster management approach in Indonesia already employs Twitter as part of the early warning system. This research suggests that, that system can be enhanced by using Twitter in a number of other ways. For example, it can be used to pre-identify individuals who can be mobilized to support the early warning and evacuation effort, to pre-identify individuals who can serve as translators during the response effort, to provide guidance on where the internet population is, to assess the extent to which they are impacted by comparing the after twitter activity to the baseline. Twitter data is also limited. For example, it cannot be used to estimate the location of the non internet users. Finally, there are possible uses, for which the current analytic technology for assessing Twitter are not sufficient; e.g., auto-identifying those tweets that are focused on disaster needs. Baseline analyses, such as that herein, are valuable in providing guidance in the use of new technologies for disaster management in ways that go beyond simple anecdotal account of their use.

7 – References

Abbott, J. P. (2011). Cacophony or empowerment? Analysing the impact of new information communication technologies and new social media in Southeast Asia. *Journal of Current Southeast Asian Affairs*, *30*(4), 3-31.

Acar, A., & Muraki, Y. (2011). Twitter for crisis communication: lessons learned from Japan's tsunami disaster. International Journal of Web Based Communities, 7(3), 392-402.

Anggunia, S. D., & Kumaralalita, L. (2014, March 13). How Indonesians Use ICT and Social Media for Disaster Management. discover.isif.asia. http://discover.isif.asia/2014/03/how-indonesians-use-ict-and-social-media-for-disaster-management/

Armstrong, C. L., & Gao, F. (2010). Now Tweet this: How news organizations use Twitter. *Electronic News*, *4*(4), 218-235.

Asur, S., & Huberman, B. A. (2010). Predicting the future with social media. arXiv:1003.5699

Barendregt, B., & Pertierra, R.l (2008). Supernatural mobile communication in the Philippines and Indonesia. In J. E. Katz (ed.), *Handbook of mobile communication studies* (pp. 377-388). Cambridge, Massachusetts: The MIT Press.

Basuki, T. (1998). Indonesia: The Web as a weapon. *Developmental Dialogue 2: The Southeast Asian media in a time of crisis* (pp. 96-103).

Birowo, M. A. (2011). Community radio and natural disasters (Indonesia). In J. D. H. Downing (Ed.), *Encyclopedia of Social Movement Media* (pp. 130-132). SAGE Publications, Inc.

Brenner, S. (1999). On the public intimacy of the new order: images of women in the popular Indonesian print media. *Indonesia*, 13-37.

Bruns, A., & Moe, H. (2013). Structural layers of communication on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), Twitter and Society (pp. 15–28). New York: Peter Lang.

Carley, K. M. (2014, forthcoming). Crisis mapping: Big data from a dynamic network analytic perspective. *Journal of Organization Design*.

Carver, L. & Turoff, M., 2007. Human-computer interaction: the human and computer as a team in emergency management information systems. Commun. ACM 50, 3 (March 2007), 33-38. DOI=http://dx.doi.org/10.1145/1226736.1226761

Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on Twitter. In Proceedings of the 20th international conference on World wide web (pp. 675-684). ACM.

Center for Excellence in Disaster Management & Humanitarian Assistance. (2011, May). Country Disaster Response Handbook: Indonesia.

Chatfield, A. T., & Brajawidagda, U. (2013, January). Twitter Early Tsunami Warning System: A Case Study in Indonesia's Natural Disaster Management. In *System sciences (HICSS), 2013 46th Hawaii international conference on* (pp. 2050-2060). IEEE.

Chatfield, A. T., Reddick, C. G., Inan, D. I., & Brajawidagda, U. (2014, June). E-government, social media, and risk perception communication at the edge of disaster: findings from the

Mt. Sinabung eruption in Indonesia. In *Proceedings of the 15th Annual International Conference on Digital Government Research* (pp. 153-162). ACM.

Chatfield, A. T., Scholl, H. J. J., & Brajawidagda, U. (2013). Tsunami early warnings via Twitter in government: Net-savvy citizens' co-production of time-critical public information services. Government Information Quarterly, 30(4), 377-386.

Cohen, R., & Ruths, D. (2013). Classifying political orientation on Twitter: It's not easy! In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (pp. 91–99). Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6128

Compton, R., Jurgens, D., & Allen, D. (2014). Geo-tagging One Hundred Million Twitter Accounts with Total Variation Minimization. arXiv:1404.7152

Couldry, N., & Curran, J. (Eds.). (2003). *Contesting media power: Alternative media in a networked world*. Rowman & Littlefield Publishers.

Cozma, R., & Chen, K.-J. (2013) What's in a Tweet? *Journalism Practice*, *7*(1), 33-46.

Dilley, M. (2005). Natural disaster hotspots: a global risk analysis (Vol. 5). World Bank Publications.

Donath, J. (2007). Signals in social supernets. Journal of Computer-Mediated Communication, 13(1), 231–251. http://doi.org/10.1111/j.1083-6101.2007.00394.x

Earle, P., Bowden, D., & Guy, M., 2011. "Twitter earthquake detection: Earthquake monitoring in a social world," *Annals of Geophysics*, 54(6): 708–715.

Eickelman, D. F., & Anderson, J. W. (Eds.). (2003). *New media in the Muslim world: The emerging public sphere*. Indiana University Press.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of lexical change in social media. -- http://arxiv.org/abs/1210.5268

Gaffney, D., & Puschmann, C. (2013). Data collection on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), Twitter and Society (pp. 55–68). New York: Peter Lang.

Gayo-Avello, D. (2012). "I wanted to predict elections with Twitter and all I got was this lousy paper": A balanced survey on election prediction using Twitter data. arXiv:1204.6441

Guha-Sapir, D., & Lechat, M. F. (1986). Information systems and needs assessment in natural disasters: an approach for better disaster relief management. Disasters, 10(3), 232-237.

Hargittai, E., & Litt, E. (2011). The tweet smell of celebrity success: Explaining variation in Twitter adoption among a diverse group of young adults. *New Media and Society*, *13*(5), 824-842.

Hecht, B., & Stephens, M. (2014). A tale of cities: Urban biases in volunteered geographic information. In Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (pp. 197–205). Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8114

Holcomb, J., Gross, K., & Mitchell, A. (2011, November 14). How mainstream media outlets use Twitter: Content analysis shows an evolving relationship. Pew Research Center Project for Excellence in Journalism and The George Washington University School of Media and Public Affairs. http://www.journalism.org/2011/11/14/how-mainstream-media-outlets-use-Twitter/

Honeycutt, C., & Herring, S. C. (2009). Beyond microblogging: Conversation and collaboration via Twitter. In Proceedings of the 47th Hawaii International Conference on System Sciences (pp. 1–10). Los Alamitos, CA, USA: IEEE Computer Society. http://doi.org/10.1109/HICSS.2009.602

Hossmann, T., Legendre, F., Carta, P., Gunningberg, P., & Rohner, C. (2011, September). Twitter in disaster mode: Opportunistic communication and distribution of sensor data in emergencies. In Proceedings of the 3rd Extreme Conference on Communication: The Amazon Expedition (p. 1). ACM.

Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. International Journal of Emergency Management, 6(3), 248-260.

Ishino, A., Odawara, S., Nanba, H., & Takezawa, T. (2012, October). Extracting transportation information and traffic problems from tweets during a disaster. In *IMMM 2012, The Second International Conference on Advances in Information Mining and Management* (pp. 91-96).

Kamada, T. & Kawai, S. (1989), "An algorithm for drawing general undirected graphs", Information Processing Letters (Elsevier) 31 (1): 7–15, doi:10.1016/0020-0190(89)90102-6.

Kireyev, K., Palen, L., & Anderson, K. (2009, December). Applications of topics models to analysis of disaster-related Twitter data. In *NIPS Workshop on Applications for Topic Models: Text and Beyond* (Vol. 1).

Kumar, S. & Morstatter, F. (2011). "TweetTracker," [Online]. Available: http://tweettracker.fulton.asu.edu.

Kumar, S., Agarwal, N., Lim, M., & Liu, H. (2009). Mapping socio-cultural dynamics in the Indonesian blogosphere. *Third International Conference on Computational Cultural Dynamics*. http://www.aaai.org/ocs/index.php/ICCCD/ICCCD09/paper/view/1019

Kumar, S., Barbier, G., Abbasi, M. A. & Liu, H. (2011). "TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief," in Proceedings of the 2011 International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 2011, pp. 661–662.

Kumar, S., Morstatter, F., & Liu, H. (2015). Twitter data analytics. Springer Briefs in Computer Science. Springer.

Kuzma, J. (2010, March). Asian Government Usage of Web 2.0 Social Media. *European Journal of ePractice*, *9*.

Landwehr, P. & Carley, K. M. (2014, forthcoming). Social Media in Disaster Relief. In W. W. Chu (ed.), *Data Mining and Knowledge Discovery for Big Data,* Vol. 1 (pp. 225–257). Berlin, Heidelberg: Springer.

Landwehr, P., Wei, W., Kowalchuck, M., & Carley, K.M. (forthcoming, this issue). Using Tweets to Support Disaster Planning, Warning and Response.

Lauder, A. (2008, July). The status and function of English in Indonesia: A review of key factors. *Makara, Sosial Humaniora*, *12*(1), 9-20. http://journal.ui.ac.id/index.php/humanities/article/viewFile/128/124

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., and Shook, E., 2013, Mapping the global Twitter heartbeat: The geography of Twitter. **First Monday**, [S.l.], apr. 2013. ISSN 13960466. Available at: <http://firstmonday.org/ojs/index.php/fm/article/view/4366/3654>. Date accessed: 15 Oct. 2015. doi:10.5210/fm.v18i5.4366.

Lim, M. (2003). The Internet and Reform in Indonesia. In N. Couldry & J. Curran (Eds.), *Contesting media power: Alternative media in a networked world* (pp. 273-288). Rowman & Littlefield Publishers, Inc.

Lim, M. (2003). The Internet, social networks, and reform in Indonesia. *Contesting Media Power: Towards a Global Comparative Perspective*, 273-88.

Lim, M. (2013). Many clicks but little sticks: Social media activism in Indonesia. *Journal of Contemporary Asia*, *43*(4), 636–657.

Mahmud, J., Nichols, J., & Drews, C. Home location identification of Twitter users. *ACM Transactions on Intelligent Systems and Technology*, 2014.

Malik, M. M., Lamba, H., Nakos, C., & Pfeffer, J. (2015). Population Bias in Geo-tagged Tweets. In Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research (pp. 18–27). Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10662

Marwick, A. E. & Boyd, D.. (2011). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. New Media & Society, 13(1), 114–133. http://doi.org/10.1177/1461444810365313

Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., & DeTar, C. (2015). Reporting, reviewing, and responding to harassment on Twitter: Women, action, and the media. Women Action Media. Retrieved from http://www.womenactionmedia.org/twitter-report/

Meier, P. 2012. “How the UN used social media in response to Typhoon Pablo,” *iRevolution* (8 December), at http://irevolution.net/2012/12/08/digital-response-typhoon-pablo/, accessed October, 2015.

Miguel, E., Gertler, P. & Levine, D. I. (2006). Does industrialization build or destroy social networks?. *Economic Development and Cultural Change*, *54*(2), 287-317.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P. & Rosenquist, J. N. (2011). Understanding the demographics of Twitter users. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 554-557).

Morstatter, F., Pfeffer, J., Liu, H. & Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter’s Streaming API with Twitter’s Firehose. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (pp. 400-408).

Muralidharan, S., Rasmussen, L., Patterson, D. & Shin, J. H. (2011). Hope for Haiti: An analysis of Facebook and Twitter usage during the earthquake relief efforts. Public Relations Review, 37(2), 175-177.

O'Connor, B., Krieger, M., & Ahn, D. (2010, May). TweetMotif: Exploratory Search and Topic Summarization for Twitter. In Proceedings of the AAAI Conference on Weblogs and Social Media.

Onishi, N. (2010, July 26). As English spreads, Indonesians fear for their language. *New York Times*, pp. A4. http://www.nytimes.com/2010/07/26/world/asia/26indo.html?pagewanted=all&_r=0

Palen, L. & Vieweg, S. (2008) The Emergence of Online Widescale Interaction in Unexpected Events: Assistance, Alliance & Retreat. CSCW, ACM Press, 117-126.

Palen, L., Anderson, K., Mark, G., Martin, J., Sicker, D., Palmer, M. & Grunwald, D. (2010) A Vision for Technology-Mediated Support for Public Participation & Assistance in Mass Emergencies & Disasters. Association of Computing Machinery and British Computing Society's 2010 Conference on Visions of Computer Science.

Palen, L., Vieweg, S., Liu, S. & Hughes, A.L. (2009) Crisis in a Networked World: Features of Computer-Mediated Communication in the April 16, 2007 Virginia Tech Event. Social Science Computer Review.

Peduzzi, P., Dao, H., Herold, C. & Mouton, F. (2009). Assessing global exposure and vulnerability towards natural hazards: the Disaster Risk Index. Natural Hazards and Earth System Science, 9(4), 1149-1159.

Philpott, S. (2000). Rethinking Indonesia: Postcolonial theory, authoritarianism and identity.

Poblete, B., Garcia, R., Mendoza, M, & Jaimes, A. (2011). Do all birds tweet the same? Characterizing Twitter around the world. In Proceedings of the 20th ACM international conference on Information and knowledge management (p. 1025). New York, NY: ACM Press. http://doi.org/10.1145/2063576.2063724

Poblete, B., Garcia, R., Mendoza, M. & Jaimes, A. (2011). Do all birds Tweet the same? Characterizing Twitter around the world. *CIKM'11*, October 24–28, 2011, Glasgow, Scotland, UK (pp. 1025-1030).

Power, R., Robinson, B., Colton, J. & M. Cameron, 2014, In Hanachi, C., Bénaben, F., and Charoy, F. (Eds), Information Systems for Crisis Response and Management in Mediterranean Countries: Proceedings of the First International Conference, ISCRAM-med 2014, Toulouse, France, October 15-17, 2014, Springer, New York.

Quarantelli, E. L. (1988). "Disaster crisis management: A summary of research findings," in Journal of Management Studies 25: 373-385.

Quarantelli, E. L. (1997). Problematical aspects of the information/communication revolution for disaster planning and research: ten non-technical issues and questions. Disaster Prevention and Management: An International Journal, 6(2), 94-106.

Rogers, R. (2013). Foreword: Debanalising Twitter: The transformation of an object of study. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), Twitter and Society (pp. ix–xxvi). New York: Peter Lang.

Ruths, D. & Pfeffer, J. (2014). Social media for large studies of behavior. Science, 346(6213), 1063–1064. http://doi.org/10.1126/science.346.6213.1063

Setiyono, B., & McLeod, R. H. (2010). Civil society organisations' contribution to the anti-corruption movement in Indonesia. *Bulletin of Indonesian Economic Studies*, *46*(3), 347-370.

Shalizi, C. R. & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, *40*(2): 211-239.

Sutton, J., Palen, L. & I. Shklovski. (2008) - Backchannels on the Front Lines: Emergent Use of Social Media in the 2007 Southern California Fires. ISCRAM.

Terpstra, T., de Vries, A., Stronkman, R. & Paradies, G. L. (2012, April). Towards a realtime Twitter analysis during crises for operational crisis management. In ISCRAM'12: Proceedings of the 9th International ISCRAM Conference.

The Asia Foundation, (2014). Elections in Indonesia. Available from https://asiafoundation.org/resources/pdfs/IndonesiaElections.pdf accessed October, 2015.

Thomas, K., Grier, C., Song, D. & Paxson, V. (2011). Suspended accounts in retrospect: An analysis of Twitter spam. In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference (p. 243). New York, NY: ACM Press. http://doi.org/10.1145/2068816.2068840

Thomas, K., McCoy, D., Grier, C., Kolcz, A. & Paxson, V. (2013). Trafficking fraudulent accounts: The role of the underground market in Twitter spam and abuse. In Proceedings of the 22Nd USENIX Conference on Security (pp. 195–210). Berkeley, CA, USA: USENIX Association. Retrieved from http://dl.acm.org/citation.cfm?id=2534766.2534784

Thomas, R. (1995), “Access and inequality”, in Heap, N., Thomas, R., Einon, G., Mason, R. and Mackay, H. (Eds), Information Technology and Society: A Reader, Sage, London, pp. 90-99.

Thomson, R., Ito, N., Suda, H., Lin, F., Liu, Y., Hayasaka, R., ... & Wang, Z. (2012, April). Trusting Tweets: The fukushima disaster and information source credibility on Twitter. In 9th ISCRAM conference (p. 10).

Tinker, T. &Vaughan, E. (2010), Risk and Crisis Communications: Best Practices for Government Agencies and Non-Profit Organizations, Booz Allen Hamilton, 2010, p. 30, http://www.boozallen.com/media/file/Riskand-Crisis-Communications-Guide.pdf.

Tobias, E. (2011). Using Twitter and other social media platforms to provide situational awareness during an incident. Journal of business continuity & emergency planning, 5(3), 208-223.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *ICWSM '14: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media* (pp.505-514). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8062

van Dijck, J. (2013). The culture of connectivity: A critical history of social media (1st ed.). Oxford ; New York: Oxford University Press.

van Oosterom, P., Zlatanova, S. & Fendel , E. (2006). Geo-information for Disaster Management, Springer, New York.

Vieweg, S., Hughes, A. L., Starbird, K. & Palen, L. (2010, April). Microblogging during two natural hazards events: what Twitter may contribute to situational awareness. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1079-1088). ACM.

Wei, W., Joseph, K., Liu, H. & Carley, K.M. (2015). The Fragility of Twitter Social Networks Against Suspended Users. In Proceedings of the ASONAM Conference.

Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C. (eds.). (2014). *Twitter and society*. New York: P. Lang.

Wong, F. M. F., Sen, S., & Chiang, M. (2012). Why watching movie tweets won't tell the whole story? arXiv:1203.4642.

Yates, D., & Paquette, S. (2011). Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake. International Journal of Information Management, 31(1), 6-13.

Zhang, D., Zhou, L., & Nunamaker Jr, J. F. (2002). A knowledge management framework for the support of decision making in humanitarian assistance/disaster relief. Knowledge and Information Systems, 4(3), 370-385.

Appendix A: Indonesia Twitter Search Specification

Keywords:

Locations:

aceh, #aceh
ambon, #ambon
bali, #bali
banten, #banten
bandar lampung, #bandarlampung
bangka belitung, #bangkabelitung
bandung, #bandung
banjarmasin, #banjarmasin
bengkulu, #bengkulu
central java, #centraljava
daerah khusus ibukota, #daerahkhususibukota
east java, #eastjava
gorontalo, #garontalo
jakarta, #jakarta
jambi, #jambi
jawa barat, #jawabarat
jawa tengah, #jawatengah
jawa timur, #jawatimur
jayapura, #jayapura
kalimantan, #kalimantan
kendari, #kendari
kepulauan, #kepulauan
kupang, #kupang
lampung, #lampung
makassar, #makassar
maluku, #maluku
mamuju, #mamuju
manado, #manado
manokwari, #manokwari
mataram, #mataram
medan, #medan
nusa tenggara, #nusatenggara
pangkal pinang, #pangkalpinang
padang, #padang
palangkaraya, #palangkaraya
palembang, #palembang
palu, #palu
papua barat, #papuabarat
pekanburu, #pekanburu
pontianak, #pontiak
riau, #riau
tanjung pinang, #tanjungpinang
tanjung selor, #tanjungselor
samarinda, #samarinda
semarang, #semarang
serang, #serang
sofifi, #sofifi
sulawesi, #sulawesi
sumatera, #sumatera
surabaya, #surabaya
west java, #westjava
yogyakarta, #yogyakarta
surakarta, #surakarta

Politics:

yudhoyono, #yudhoyono
boediono, #boedino
bakrie, #bakri
sukarnoputri, #sukarnoputri
iskandar, #iskandar
matta, #matta
rajasa, #rajasa
wiranto, #wiranto
suhardi, #suhardi
demokrat, #demokrat
golkar, #golkar
pdip, #pdip
pdi-p, #pdi-p
pkb, #pkb
pks, #pks
hanura, #hanura
gerindra, #gerindra
pancasila,#pancasila

Indonesian language:

kepo	bokop
cowok	nyokap
cewek	

Related to Indonesia:

sunda	bahasa	
sundanese	bahasa	indonesia

Related to disaster:

tsunami	butuh pengobatan
makanan	butuh obat-obatan
bekal	perlu obat
lauk	obat
makan	obat-obatan
air	anak hilang
minuman	food
halte	water
tempat berlindung	shelter
orang hilang	missing person
korupsi	corruption
penipuan	fraud
butuh makanan	need food
perlu makanan	need water
perlu makan	need shelter
perlu bekal	need medicine
butuh minuman	medicine
butuh tempat berlindung	missing child
butuh tempat bernaung	

Bounding Geobox:

Southeast Corner: (-9.5, 95.0)
Northwest Corner: (6.0, 141.0)

User IDs:

42420346	47363197	28302919	1056601446
218437274	94339403	62020959	548489130
33873560	42513459	31178630	101420781
69183155	62469487	168082298	142209668
128841198	57261519	1962874675	77957047

Appendix B: Padang Twitter Search Specification

Keywords:

Locations:

padang, #padang
bungus teluk kabung, #bungustelukkabung
koto tangah, #kototangah
kuranji, #kuranji
lubuk begalung, #lubukbegalung
lubuk kilangan, #lubukkilangan
nanggalo, #nanggalo
padang barat, #padangbarat
padang selatan, #padangselatan
padang timur, #padangtimur
padang utarah, #padang utarah
pauh, #pauh
minangkabau, #minangkabau
andalas university, #andalas
fauzi bahar, #fauzibahar
emmahaven, #emmahaven
teluk bayur, #telukbayur
bayur bay, #bayurbay

Politics:

yudhoyono, #yudhoyono
boediono, #boedino
bakrie, #bakri
sukarnoputri, #sukarnoputri
iskandar, #iskandar
matta, #matta
rajasa, #rajasa
wiranto, #wiranto
suhardi, #suhardi
demokrat, #demokrat
golkar, #golkar
pdip, #pdip
pdi-p, #pdi-p
pkb, #pkb
pks, #pks
hanura, #hanura
gerindra, #gerindra
pancasila, #pancasila

Related to Disaster:

tsunami
makanan
bekal
lauk
makan
air
minuman
halte
tempat berlindung
orang hilang
korupsi
penipuan
butuh makanan
perlu makanan
perlu makan
perlu bekal
butuh minuman
butuh tempat berlindung
butuh tempat bernaung
butuh pengobatan
butuh obat-obatan
perlu obat
obat
obat-obatan
anak hilang
food
water
shelter
missing person
corruption
fraud
need food
need water
need shelter
need medicine
medicine
missing child

Bounding Geobox:

Southeast Corner: (-1.05, 100.25)
Northwest Corner: (-0.75, 100.5)

Institute for Software Research • Carnegie Mellon University • 5000 Forbes Avenue • Pittsburgh, PA 15213-3890