Final Paper: Exploring the Role of Geographic Diffusion in Informational Cascades

Devin Gaffney

25 April 2015

Introduction

That any particular information and communication technology (ICT) carries with it the potential for collapsing geographic distance between individuals is not a particularly novel idea – Cairncross' (2001) "The Death of Distance" is particularly instructive in this right. In an article discussing the work, Cairncross states that "[w]e are beginning to learn that we have often more in common with people on the other side of the world than with the folks who live next door" (Cairncross, 2002). While it's certainly easy to *imagine* how the Internet and ICT's generally may bring with it the ability to collapse geography dramatically, it's another thing altogether for an ICT to *actually* transcend geography. In fact, much evidence has been raised since the publication of Cairncross' book that leads to the conclusion that social ties remain relatively geographically bounded (Scellato, Noulas, Lambiotte, & Mascolo, 2011).

Beyond refuting Cairncross' work, it is important to be generally reminded of how robust geographic homophily remains – McPherson, Smith-Lovin, and Cook go so far as to call it "[p]erhaps the most basic source of homophily". In other terms, despite the Internet's potential transformative quality in collapsing geographic distance, the tendency for geographically proximate individuals to be socially tied remains to be robust, and has been known to be an important factor for quite some time (Zipf, 1949). In the context of online social networks, one may then reasonably expect a degree of clustering between individuals within a constrained geographic space. Other forms of homophily

may exist for individuals *within* that space – indeed, McPherson et al. also points toward familial, organizational, and isomorphic homophilic tendencies. At a high level then, in the aggregate, one may envision online social networks as homophilic clusters according to attributes such as geography, family ties, professional and interest organizations, all with various overlaps according to social ties falling outside these general tendencies. The homophily phenomenon is both a source of great interest as well as great methodological concern in the study of online social networks, and is one of two focuses in this work.

The other phenomenon of interest is the phenomenon of informational cascades. In both academic and professional contexts, the concept of "going viral" has been of considerable interest in terms of both explicating the mechanisms underlying it as well as predicting the conditions required for it to occur (Kamath, Caverlee, Lee, & Cheng, 2013; Galuba, Aberer, Chakraborty, Despotovic, & Kellerer, 2010; Bakshy, Hofman, Mason, & Watts, 2011; Baños, Borge-Holthoefer, & Moreno, 2013; Lerman, Ghosh, & Surachawala, 2012). This phenomenon is only loosely defined thus far – it has been described as "going viral" and an informational cascade, among other terms (Nahon & Hemsley, 2013; Cheng, Adamic, Dow, Kleinberg, & Leskovec, 2014). While many papers have explored the dynamics of information cascades, few have provided even a cursory definition of the phenomenon (Cha, Haddadi, Benevenuto, & Gummadi, 2010; Rattanaritnont, Toyoda, & Kitsuregawa, 2012; Gayo-Avello, Peter Gloor, Castillo, Mendoza, & Poblete, 2013; Cha, Benevenuto, Ahn, & Gummadi, 2012). Cheng et al. provides a relatively loose definition which still provides at least a sense of the phenomenon of interest: "In certain instances, a photo, link, or other piece of information may get *reshared* multiple times: a user shares the content with her set of friends, several of these friends share it with their respective sets of friends, and a *cascade* of resharing can develop, potentially reaching a large number of people". In other terms, it is relatively common for content to be reshared a few times – an information cascade is distinct when the number of reshares some particular content receives reaches a point at which it rapidly outpaces the typical patterns most content experiences.

It is the goal of this work to explore how these two seemingly disparate phenomena may be related. First, it will be hypothesized that if geography is, in the aggregate, a primary factor for homophilic tendencies, then actors within non-geographic based clusters of individuals, e.g. interest-

based clusters, may span across more distinct clusters in turn. By spanning across more distinct clusters in turn, content that emerges from non-geographically bounded clusters of individuals may have a wider initial geographic audience, which may in turn inhibit that content's ability to become an informational cascade. Ultimately, the relationship between the initial geographic dispersion of a particular piece of content and the degree of popularity for a particular piece of content is explored. While other factors clearly exist, such as the nature of the content itself, it is the goal of this work to argue that initial network structural topography, specifically in terms of geographic dispersion, play a clear role in determining what content is ultimately involved in a viral cascade.

Methodology

In attempting to explicate the relationship between initial geographic dispersion of individuals involved in an informational cascade and an informational cascade's ultimate popularity within an online social network, several methodological concerns must be raised and resolved. First, what does "initial geographic dispersion" mean in the context of an informational cascade? Second, what online social network[s] should be considered, and what informational cascades should be considered as examples? Third, how can one detect geographic position of individuals? Finally, what confounding factors complicate understanding the relationship this paper seeks to understand?

Defining "Initial" Growth

Informational cascades take place within the context of a connected cluster resharing the content – in the context of most online social networks, information appears on a timeline of some form from people that an individual follows or is friends with (boyd, Golder, & Lotan, 2010). The crux of the argument in this work is that the initial geographic dispersion within the cluster of actors from which a piece of content emerges is related to the ultimately popularity of an informational cascade. Here, one can easily operationalize popularity as either the total number of "reshares" of a piece of content or the number of unique individuals who "reshared" the content – ultimately, some quantitative measure of the magnitude of an informational cascade is the salient dependent variable.

The concept of geographic dispersion is also relatively straightforward. What is sought after is a measure of geographic distance between all individuals in a group – one could easily operationalize this as the sum of euclidean distances between each individual in the group. Another measure could instead be the sum of euclidean distances between each interacting individual within the group, which would likely more accurately capture the distance upon which communication actually occurred as opposed to the total distance of the group generally. Of course, there are questions of geolocation, which will be raised later, but the point remains that this metric seems to be operationalized relatively clearly.

The last point which remains to be operationalized is the definition of "initial". At some very early point, an informational cascade is indistinguishable from the small set of reshares that categorize the vast majority of unnoticed informational non-cascades to the point that they are not of interest in this work. Of course, at the other end, an informational cascade has grown to the point that the mechanism of interest, the rapid acceleration of an informational cascade, has long past. Turning to the network literature, some work provides a relatively straightforward option.

In the initial stage of any informational cascade, one actor in an online social network creates the content. Representing this as a network, one could envision this as a network consisting of only one node. As another actor reshares the content from this initial actor, a possible representation of this interaction would be a network of two nodes (representing actors) and one directed edge (representing the reshare). As the informational cascade continues, actors continually reshare content from others, and the network grows. While the literature considers slightly separate networks (specifically, Erdős-Renyi networks), an interesting property emerges when the average degree of nodes in the network approaches 1. Specifically, when the average degree of nodes is less than one, the graph is said to be in the "subcritical regime" (Albert & Barabási, 2002; Erdős & Rényi, 1959). In Albert and Barabási work, the sub-critical regime is typified by trees of actors within the network, disconnected from one another. Once the average degree surpasses 1, the graph enters the "critical regime", the giant component of the graph appears, and a fully connected network quickly stitches together previously unconnected smaller subgraphs.

In terms of online social networks, and specifically, actors engaged in informational cascades, the concept of the critical regime is substantively useful – at this point in an informational cascade,

one may envision the emergence of a giant component the point at which a hitherto dispersed set of individuals engaging with the content begin to find themselves in a coherent conversation – what was once isolated groups engaging in the cascade becomes one large conversation about the content. Indeed, a review of literature concerning online social networks readily adopts the notion of the emergence or presence of a giant component as a substantively meaningful point at which a conversation online becomes coherent (Macskassy, 2012; Nagar, Seth, & Joshi, 2012; Ardon et al., 2013; Aragón, Kappler, Kaltenbrunner, Laniado, & Volkovich, 2013). Given that previous literature employs this point as substantively useful, and given that there is a clear point at which the emergence of a giant component begins to appear, it is reasonable to define initial actors as actors who were reshared content in the context of an informational cascade immediately before the giant component appears, or before the average degree of actors within the network equals 1. Of course, some networks in practice may never surpass an average degree of 1 – in these cases, all actors should be considered owing to any practical limitations. Additionally, networks may surpass an average degree of 1 and then fall once again – in these cases, the actors present before the first point at which the degree *surpasses* 1 would likely be a reasonable restriction in defining "initial" actors, which in turn defines "initial geographic dispersion".

Of the 300 terms that were collected (which will be discussed below), in only 190 cases did $< k >$ exceeded 1 during the two week sample from from March 1 to March 15. While $< k >> 1$ is theoretically valuable, the relative dearth of cases that exceed this value leads towards adding in additional alternate approximations of the same phenomenon. Specifically, two other loose working definitions of "initial" come to mind. First, the logic behind the $< k >= 1$ lies in the emergence of a single dominant connected component for any given network. While $< k >= 1$ is an approximation for this phenomenon, a threshold on the size of the largest connected component as a proportion of the total size of the network in terms of nodes could also be used. In this case, the paper also provides an alternative model for "initital" geographic dispersion as the point at which $log(n)/log(N) > 0.5$ is first satisfied, where n is the number of nodes in the giant component and N is the total nodes in the network. Additionally, a third temporal threshold is added to explore if structure-independent definitions of "initial" activity may provide further insight. Operationalized, the temporal threshold is set to include all original actors involved with a particular topic if they tweeted within 6 hours

of the first tweet about the topic.

Cascade Selection

One of the difficulties in studying informational cascades rests in the fact that many cascades fail to ever become large. In other terms, the most notable cascades, those which transit through a network widely, are likely the exception to the much more frequent cases of very limited informational cascades (Goel, Watts, & Goldstein, 2012). For this reason, as well as problems of general selection bias, selecting from a subset of large, well known informational cascades would only afford for a biased look at any statistical inference for informational cascades generally. At the same time, however, by casting too broad a net, the vast majority of informational cascades that never ultimately grow beyond only a few participants would likely yield little in the way of exploring the medium and larger cascades that this paper seeks to explore.

One particular online social network provides a data source that may potentially find a middle ground between these two tensions. Twitter's "trending topic" API endpoint provides a way to collect cases of informational cascades without introducing much inherent researcher selection bias. Of course, trending topics in and of themselves carry their own biases – they are, after all, selected by an opaque algorithm which likely has it's own definition of popularity (Kwak, Lee, Park, & Moon, 2010). Still, content that appears as a trending topic is used widely in academic research without much methodological concern – in fact, efforts to reverse engineer the algorithm, as well as statements from Twitter itself, have helped to demystify the algorithm behind trending topics (Nikolov & Shah, 2012; Twitter, 2010; Lotan, 2011).

Additionally, given the loose definition of an informational cascade, there is ample room to argue that trending topics, regardless of the underlying algorithm propagating them, are a function of novelty and volume around a given term that is being reshared, or in context, retweeted within the network. In other words, trending topics are indicative of at least one class of informational cascade. Furthermore, trending topics only become problematic in terms of arguing for any organic process once the topic becomes trending – by focusing on content created before the topic becomes a trending topic, it is reasonable to argue that negative effects of trending topics, notably the high

occurrence of spammers, can be largely avoided (Benevenuto, Magno, Rodrigues, & Almeida, 2010; Yardi, Romero, Schoenebeck, et al., 2009).

Ultimately, trending topics are problematic, but they do provide a clear path towards selecting cases of informational cascades that go beyond the trivially small but ensure that the selected cases of informational cascades are not only the largest and most exceptional cases. This is further bolstered by the localization of trending topics – while topics used to trend only globally, a total of 466 locales, ranging from city-level to national-sized boundaries, have been established to track trending topics, which further aids in ensuring that trending topic selection provides a broad set of diverse informational cascades from which to select a set of cases.

Geographic Detection

One lesser consideration is operationalizing the geographic placement of actors. As argued in the section above, Twitter is an ideal environment for observing informational cascades due to the presence of the trending topics API endpoint – in the context of Twitter, much work has been done concerning Twitter's geospatial attributes. How may one derive the actual location of actors within the context of Twitter? Takhteyev, Gruzd, and Wellman opted to hand-code actors in their work, as the number of actors was a tractable size for a team of researchers. In the context of this question, where a set of informational cascades must be considered, there will likely be many non overlapping sets of actors ultimately leading to a large number of actors beyond a point where hand-coding can feasibly be conducted. In the absence of a hand-coded set, a geocoding API, or a web service that converts a string of text into a structured latitude and longitude representation (Teske, 2014). Indeed, previous related work has employed such geocoding APIs (Java, Song, Finin, & Tseng, 2007).

Looking towards automated geocoding, current literature is not particularly promising. Graham, Hale, and Gaffney cautions against full faith in geocoding APIs alone. Specifically, their work collected all tweets containing GPS data over one month, and then measured the degree to which their GPS-level data differed from the stated location for the actor's Twitter profiles. In this work, the authors found large discrepancies in terms of how well profile locations were able to predict

GPS locations. Two factors mitigate findings from Graham et al.'s work. First, data collection for this research occurred in a time when GPS-level data was relatively nascent on the platform – indeed, the authors cite the relative dearth of GPS data as potentially problematic. Second, while geocoding APIs may be inaccurate in a large subset of cases, if the subset of cases is not systematically biased towards any particular locational information, the results from the geocoding API should be equally inaccurate for any particular location, which would introduce a large amount of noise for geocoding actors in informational cascades, but equally so.

In other words, while geocoding APIs may be inaccurate, this inaccuracy may not result in any statistically substantive issue as the errors may be evenly dispersed. These questions of algorithmic opacity, ground truth as far as geographic location is concerned, and data analysis at scale are in many ways tied to the issues raised by boyd and Crawford. Ultimately, geographic detection may be problematic, but the only tractable attempt to define the precise geographic position of alters given the constraints and scale of this question lies upon the ability to successfully employ geocoding APIs – future work should make efforts to ensure algorithmic accuracy through a representative sampling of human coding as a complement to the geocoding attempts. Additionally by using results from several geocoding APIs, and considering their relative similarity in terms of how their results relate to the dependent variable, will add at least some degree of robustness to the findings.

Confounding Factors

To be clear, it is unlikely that initial geographic dispersion is the sole factor that influences the degree to which content approaches an informational cascade. Previous work has explicitly concerned itself with non-geographic parameters, instead option to analyze how factors such as the degree of the initial actor impacts an informational cascades ultimate popularity (Cha et al., 2010). Additionally, early work on the subject considered attributes such as the age of the account, the total number of posts per account, as well as the indegree and outdegree of nodes as potentially predictive metrics (Bakshy et al., 2011). Importantly, previous work also acknowledges that while the topological properties of individuals is an important explanatory variable, "while large follower count and past success are likely necessary features for future success, they are far from sufficient" (Bakshy et

al., 2011). Still others focus on the initial topographical properties of an information cascade, which echoes a similar sentiment in work concerning the wide spread of information cascades across communities in the initial stages being a predictor of future growth (Cheng et al., 2014; Weng, Menczer, & Ahn, 2013).

For this reason, a complete analysis should include some of these network metrics in order to assess their relative utility. Complicating the matter, however, is that user-level and tweet-level attributes are a different observational level than the one currently under examination – while many studies have considered the initial actor instigating an information cascade, fewer consider the initial set of actors collectively. Operationalizing this problem would involved constructing some composite score of individual attributes, such as the sum, median, or mean of indegree (or follower counts in the context of Twitter). One may also use the values of the initial actor in each informational cascade as an additional potential explanatory variable. Including some subset of these variables is important in order to both respond to previous literature as well as determine the degree to which initial geographic dispersion is substantively valuable.

Operationalization

Having considered several important methodological aspects of this study, a clear design can now be offered: first, a sample of trending topics from Twitter's trending topics API endpoint will be collected across all 466 locales over the course of several days. From this set, a sample of trending topics will be randomly selected – for each randomly selected topic, content from a long-standing "gardenhose" (or a random 10% sample of all of Twitter's content) will be collected from several days prior to the initial discovery of the trending topic (Conover et al., 2011). Then, a network graph will be constructed for each trending topic, where nodes represent users and edges represent mentions and retweets between users. For each definition of "initial" as discussed above, all users within the network at that point will be deemed as the individuals whose geographic dispersion is of interest. For each of these trending topics, several variables will be collected.

First, the total number of distinct nodes and edges within the full two weeks from the inception of an informational cascade can be held to operationalize their popularity. Of course, some of the

trending topics within the sample will not be entirely novel to the Twitter network. Regardless of if the trending topic has already been popular (e.g. the Summer Olympics every four years, New Years Eve annually, and perennial news stories), any particular iteration of an information cascade is assumed to be a novel case in and of itself. Further, defining the degree to which something is entirely novel to the entire Twitter network is methodologically intractable given current research tools. The two week-long range, while artificial, should provide a long enough scope to capture the bulk of the informational cascade – indeed, Asur, Huberman, Szabo, and Wang note that within 48 hours much of the informational cascade has already occurred.

Second, "geographic dispersion" is operationalized all geocodable distances between interacting individuals within the initial stage of the network. In this way, geographic dispersion can be measured as a function of the actual interactions that occurred. As was noted above, other metrics should also be included in order to assess the relative impact that initial geographic dispersion has on the ultimate size of an informational cascade. For this reason, the average number of followers, as well as the average number of tweets are included as potential confounding variables. While including the first participant may be substantively useful, the data source cannot provide this, as it is a random 10% sample of all content from Twitter. As a result, the original instigator may not be present – while the data could be included, many more methodological questions, not the least of which involves defining who qualifies as the original instigator, are involved and thus this data is omitted.

Data Review

From March 4 to March 9, a connection was established to the Twitter API that consistently polled for all updated trending topics across Twitter's 466 topic locales. This resulted in a set of 2,691,989 unique appearances of trending topics, with 14,589 globally unique terms. In addition to organic trending topics, Twitter has also monetized this feature, allowing organizations to "purchase" the space generally allotted for trending topics, allowing them to place their own term instead (ostensibly to drive traffic to that topic through the high visibility of the trending topics feature) (Twitter, 2014). Of the entire set of terms, 2,686,229, or roughly 99.7%, of all trending topics in the set were

organic trends. In order to ensure that topics were relatively novel and were not caught partly through any pre-existing informational cascade, terms were randomly selected from the set only if they had not appeared in any locale before March 7. This selection resulted in a total of 6,544 unique topics. From this set, 300 initial topics were selected for analysis. While the total sample size for this work is relatively small, the scale of work required to analyze each network is considerable, so a smaller sample is more feasibly tractable.

A 10% random sample of all of Twitter ranging from mid 2011 to the current time (April 2015) was made available for review. For each of the 300 trending topics, tweets were extracted from this archive if they matched the term and occurred between March 1, at midnight UTC and March 15, at Midnight UTC. For each of the trending topics, for each hour between March 1 and March 15, topics were analyzed for any new tweets that would add additional nodes and edges to each graph, and the average degree (number of edges / number of nodes) is recorded along with the current size of the giant component and the total number of nodes.

Findings

Of the 300 trending topics collected, 190 networks experienced a point at which $< k >= 1$, 293 networks experienced a point at which $log(n)/log(N) > 0.5$, and 280 networks existed for 6 or more hours. For each of these cases, if the network failed to reach the threshold, the entirety of tweets was considered the initial set – further work should examine if instead, these cases should be omitted. These cases remain in this work however – as they never reached their respective threshold (eg, the phenomenon only lasted 5 hours or the network only had a maximum $< k >$ of 0.85), they strictly still fit the original definition of "initial" as discussed above.

Of the various geocoding utilities available, this work leverages geocoding APIs that are relatively free of rate limits due to the time constraints of geocoding large numbers of individuals at a large scale as is required for this work. Indeed, the total number of tweets under study, across the entire set of 300 trending topics, is 13,079,420 – even with some overlap, geocoding this amount is a large task. Of the Geocoding APIs available, ESRI, Bing, and Nominatim (also known as Open Street Maps, or OSM) provide APIs without discernible rate limits. For each trending topic, and for each

Geocoding API, and for each definition of "initial" as discussed above, the distance between each edgewise interaction was calculated, if in fact both nodes for an edge could be successfully geocoded.

If the calculation was successful in terms of returning a valid geolocated point for both the interactants, the distance was added to a distribution of distances for the entire network – misses were also recorded, but distances were omitted in those cases. This distribution, in turn, was then analyzed in terms of summary statistics which then allow for further analysis. This entire process is in fact ongoing, as more data still can be calculated. A small version of existing data is offered here, and table 1 lays out the current sample's size, while table 2 displays the number of cases in which there are more than zero total edgewise geolocations that were successful. An important note must also be made: data is still forthcoming in terms of the completeness of geocoding the distance of each edge. While the results here represent a significant subset of all data, further data may result in further refinements of the results.

Geocoder	$< k >> 1$	$log(n)/log(N) > 0.5$	First Six Hours	Total
Bing	182	194	194	570
ESRI	182	194	194	570
Nominatim (OSM)	183	194	194	571
Total	547	582	582	1711

Table 1: Case counts per geocoding service and "initial" threshold definition

Geocoder	$< k >> 1$	$log(n)/log(N) > 0.5$	First Six Hours	Total
Bing	114	69	36	219
ESRI	114	69	36	219
Nominatim (OSM)	115	68	36	219
Total	343	206	108	657

Table 2: Case counts per geocoding service and "initial" threshold definition with successful geocodes present

Typically, many distributions surrounding Twitter data is heavily power law distributed. This dataset is not different in that regard. Figure 1 shows the significant skewedness of the total number of tweets in each of the 300 topics. Converting it to a log space, however, makes the data relatively normal distributed as is shown in figure 2. Similarly, the data surrounding the average and standard deviation of geographic distances for edges per topic are skewed – figures 3 and 4 reflect this, while the log of each of these, figures 5 and 6 provide a view into a dataset that shows essentially two

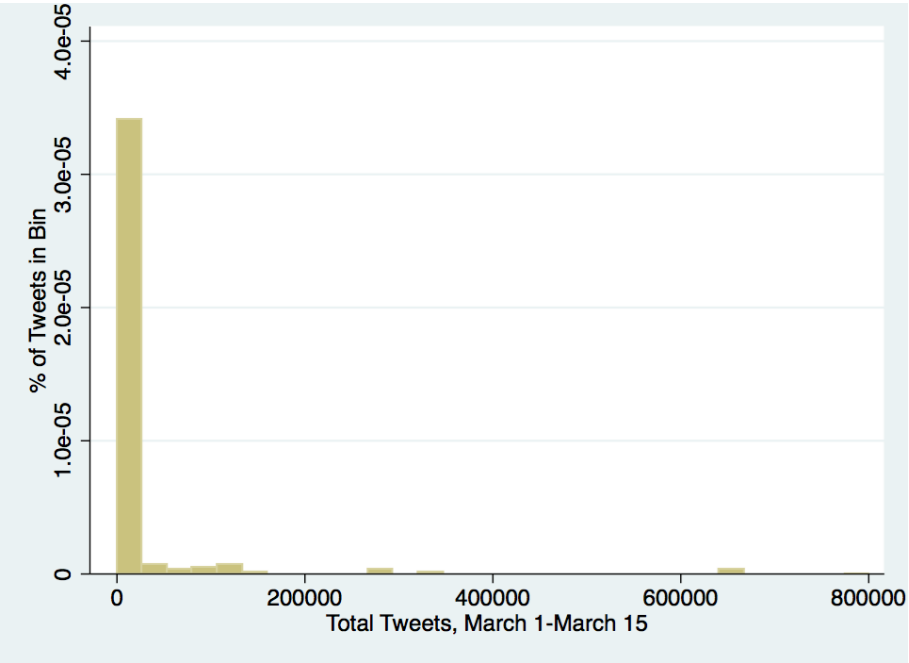

Figure 1: Distribution of total tweets for all trending topics within dataset between March 1 and March 15 (194 total cases present). Note that this data is highly skewed, with many topics reaching only a small level of popularity.

distinct distributions in both cases – in one, we see high zero counts, and in the other, we see a roughly normal distribution of distances further out. This will become clearer as the analysis continues.

Several analytical possibilities exist. First, the simplest way to explore any possible relationship here is to plot the log of the total number of tweets per topic as a function of the log of the mean and log of the standard deviation of geographic distance between measured edges within the given topic. Figures 7 and 8, respectively, plot those results. As is clear, the phenomenon earlier alluded to, wherein there are high zero counts and then a relatively normal distribution in the histograms featured in figures 5 and 6, leads to a clear insight: there are two types of trending topics.

In one case, where the mean and standard deviation of distances are zero, are localized trending topics. In other terms, low or zero distance cases are cases in which geocoders identify individuals living in the same locality. This makes substantive sense as well – typically, individuals who include geocodable data in their user bios list general localities, such as the name of their major metropolitan area, rather than specific addresses – thus, geocoders will likely geocode "Portland", "Portland, Oregon", and "Portland, OR" identically, which then resolves to a net distance between interactants

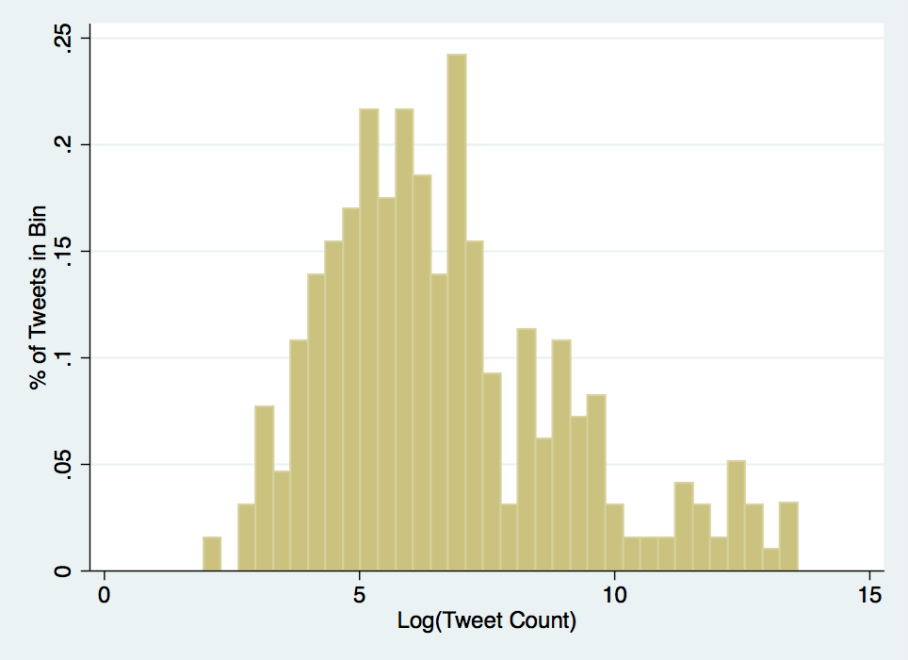

Figure 2: Distribution of the log of total tweets for all trending topics within dataset between March 1 and March 15 (194 total cases present). Note that this data is highly skewed, with many topics reaching only a small level of popularity. The log of this data, however, yields a relatively normal distribution

as zero. In other cases, there's a different result – actors interact at a distance. This is the other distribution being witnessed in this data.

The primary hypothesis of this work surrounds the notion that trending topics can be predicted by the initial geographic dispersion of their interactants. Thus far, it has become clear that some category of trending topics are local, while others are not local. In the interest of fully answering the question this paper puts forth, the cases where trending topics are purely local are no longer of interest. While they clearly exist, the goal is to show that initial geographic dispersion is positively correlated with ultimately popularity – while this is clearly not the case for a subset of local cases, can something be found for the other cases where geography does in fact change? Two approaches are considered: first, on average, how geographically far are two interactants away from one another, and second, on average, how geographically varied the average interactants are away from one another. Operationalized, the former can be represented by the mean distance per topic, and the latter can be represented by the standard deviation. The logic behind including the standard deviation is that while the average may converge around a certain type of case (e.g. two major cities being involved), the standard deviation helps in accentuating the cases where there are fewer edges

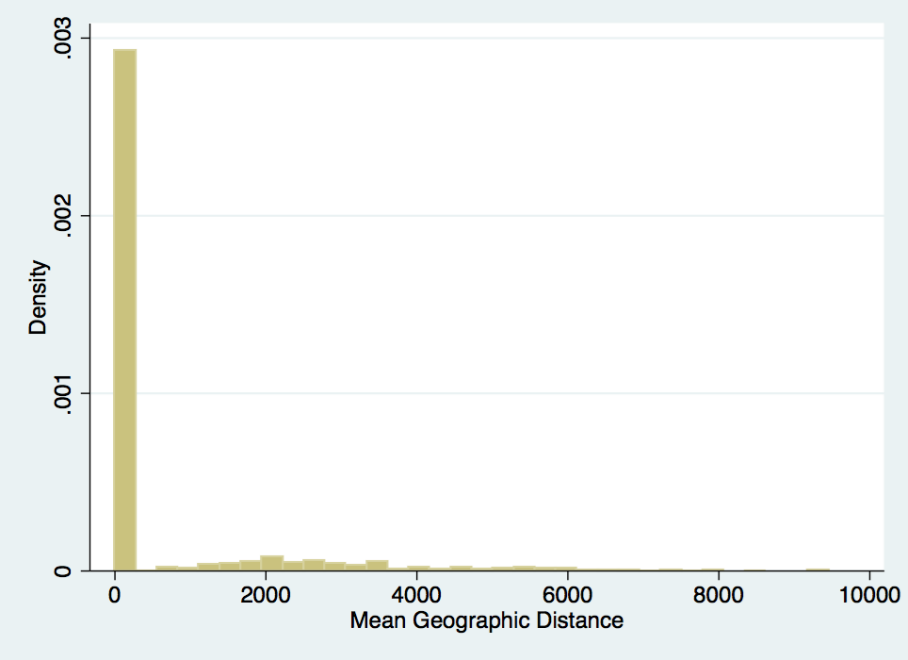

Figure 3: Distribution of mean distances for all edges geocoded per topic. Note that this data is highly skewed.

but across more varied distances with regard to the mean. Table 3 shows R^2 values from a simple linear correlation predicting tweet count by mean and tweet count by standard deviation separated by both Geocoder and "initial" definition, while table 4 reports Spearman's ρ as a non-parametric estimate, as both the mean as well as the standard deviation are not normally distributed.

Log(Standard Deviation)			
Geocoder	$< k >> 1$	$log(n)/log(N) > 0.5$	First Six Hours
Bing	**0.0328**	**0.3305***	0.0070
ESRI	**0.0317**	**0.3263***	0.0324
Nominatim (OSM)	**0.0469***	**0.2959***	0.0324
Log(Mean)			
Geocoder	$< k >> 1$	$log(n)/log(N) > 0.5$	First Six Hours
Bing	0.0076	**0.0743***	0.0001
ESRI	0.0093	**0.0419**	0.0092
Nominatim (OSM)	0.0119	0.0297	0.0140

Table 3: Regression results per geocoder and "initital" definition. $p < 0.05$ is reported in bold with star, $p < 0.10$ is reported in bold.

As tables 3 and 4 suggest, different definitions of "initial" matter. In the case of this work, defining the initial stage of a trending topic purely in terms of it's temporal nature provides little insight. Indeed, whether one accepts a linear regression or a non parametric test as a better-fitted

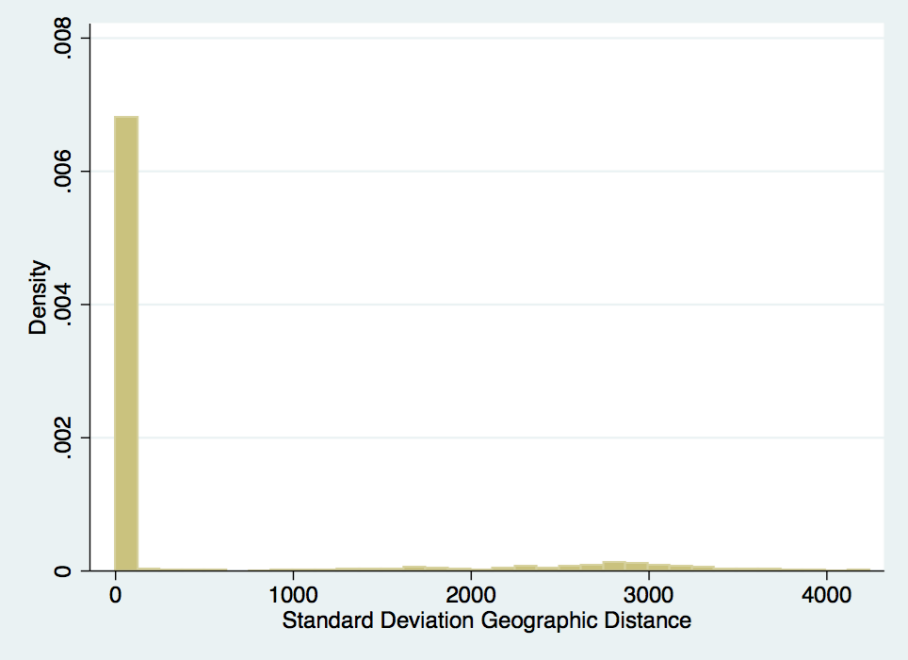

Figure 4: Distribution of standard deviation of distances for all edges geocoded per topic. Note that this data is highly skewed.

Log(Standard Deviation)			
Geocoder	$< k >> 1$	$log(n)/log(N) > 0.5$	First Six Hours
Bing	**0.2993***	**0.5392***	0.0425
ESRI	**0.2838***	**0.5576***	0.1192
Nominatim (OSM)	**0.4976***	**0.2959***	0.1312
Log(Mean)			
Geocoder	$< k >> 1$	$log(n)/log(N) > 0.5$	First Six Hours
Bing	0.0145	-0.0278	-0.0695
ESRI	0.0274	-0.0863	0.0776
Nominatim (OSM)	0.0301	-0.1380	0.1513

Table 4: Spearman's *rho* per geocoder and "initital" definition. $p < 0.05$ is reported in bold with star, $p < 0.10$ is reported in bold.

statistic, the first six hours of a trending topic with regards to the measured geographic dispersion of participants has very little bearing on the ultimate popularity of the topic. Meanwhile, $< k >> 1$, while theoretically valuable, is, as discussed previously, a crude but easily calculable proxy for the actual variable of interest, $log(n)/log(N)$. In the scope of this work, the threshold for $log(n)/log(N)$ is set at 0.5, which results in surprisingly robust results in both statistical tests with regards to the log standard deviation. In substantive terms, these tests help towards making a case that early widely varied geographic spread of interactants discussing a particular topic may indeed contribute

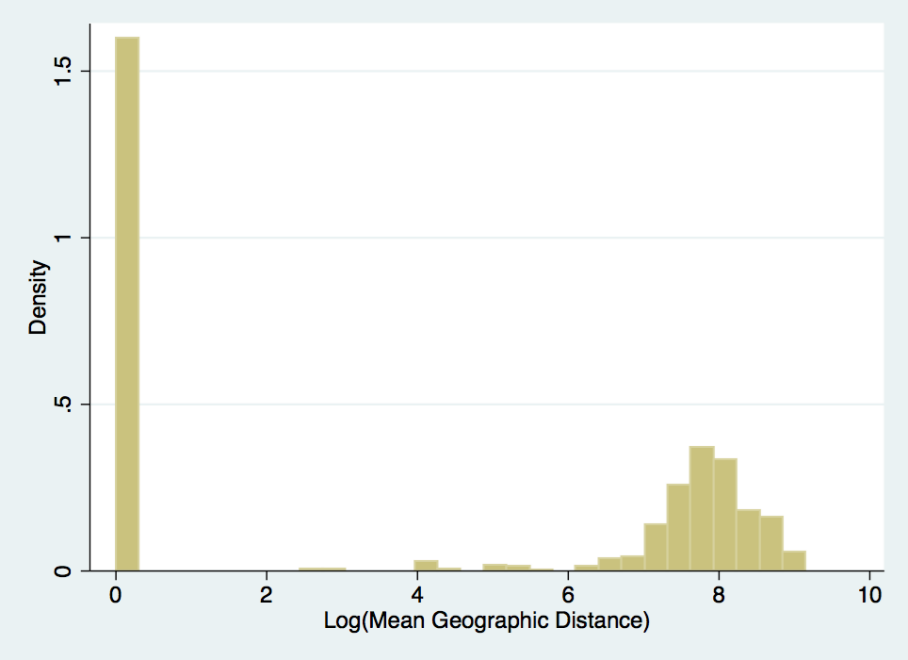

Figure 5: Distribution of mean distances for all edges geocoded per topic. Note that this data is highly skewed.

significantly to that topics ultimate popularity when cases of localized trending topics are omitted.

Discussion

The original hypothesis of this paper submitted that initial geographic dispersion of interactants within a given topic may predict the ultimate popularity of the given topic in an online social network. This hypothesis is based off of a simple inference that can be drawn from McPherson et al.'s seminal work: homophily is an organizing principle of social networks – of the different forms of homophily that exist, geographic homophily, regardless of the declarations of some, remains a primary organizing principle even in the case of online social networks. Therefore, non-geographic homophilic networks typically have an advantageous topographical position in spreading a topic, as each constituent actor typically exists within the context of their own distinct geographically homophilic community, and as a result, can spread a topic to a wider set of clusters of individuals more efficiently.

This paper aimed to explore this hypothesis by taking a sample of trending topics from Twitter in order to explore how initial geographic dispersion factored into the ultimate popularity of a

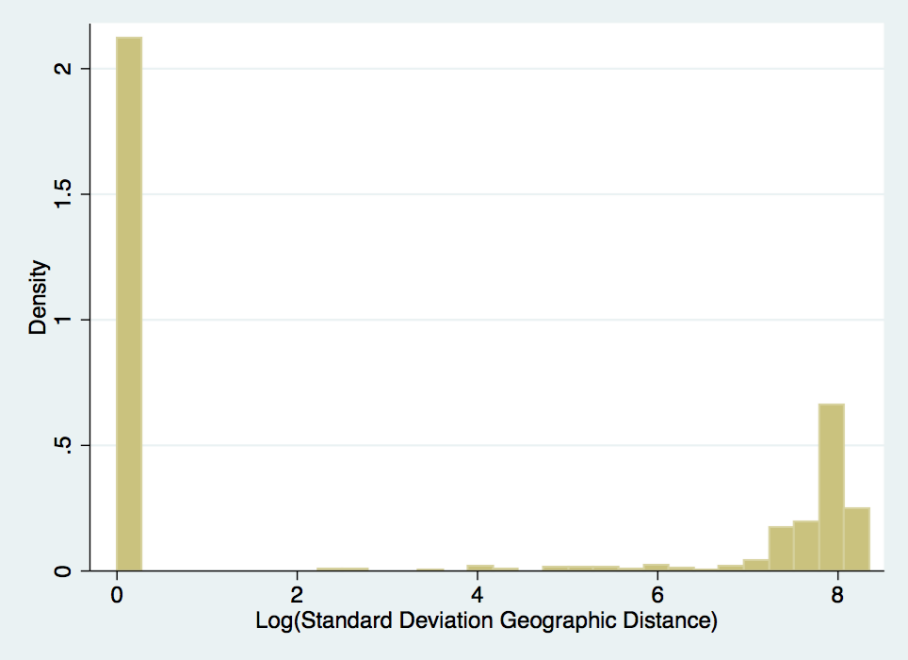

Figure 6: Distribution of standard deviation of distances for all edges geocoded per topic. Note that this data is highly skewed.

topic. "Initial geographic dispersion" was operationalized in two separate phases. "Initial" was operationalized as either the point at which $< k >$ first exceeds one, $log(n)/log(N)$ exceeds 0.5, or the first six measured hours from the occurrence of the first tweet mentioning a topic from several days before the topic trended. "Geographic dispersion" was operationalized as the distribution of distances between interactants at each operationalization of "initial". Summary statistics about the distributions of "geographic dispersion" were drawn from each of these definitions for each network at the point that they met these definitions, and those summary statistics were measured against the ultimate number of tweets for each trending topic. Findings suggest that there exists no relationship between "initial" as defined by the first six hours, a marginal relationship for $< k >$ as the definition of "initial", and a moderate to strong relationship for $log(n)/log(N)$ as the definition of "initial", which, indeed, was the true variable that $< k >$ sought to approximate.

As has been shown, the hypothesis is only partially supported by the evidence. There are a clear subset of cases where topics stay localized. One may imagine these as a subset of cases rife throughout this dataset. Collected in early March of 2015, the data bears several cases of "march madness" related topics, topics which are likely to be highly localized, which has previously been identified as trend in current literature (Zubiaga, Spina, Fresno, & Martínez, 2011; Kwak et

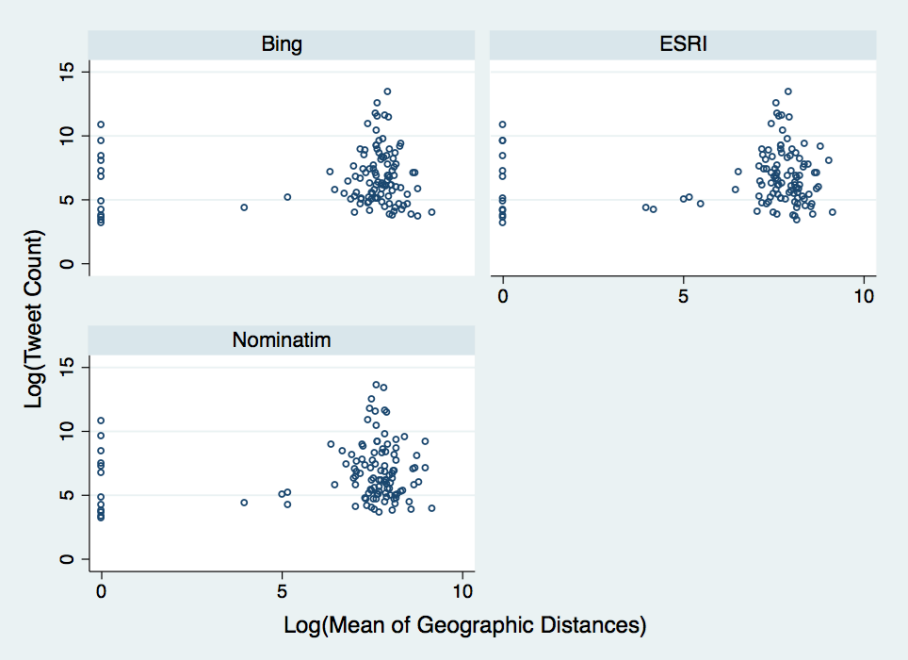

Figure 7: Scatterplot of log mean of distances for all edges geocoded per topic by the log of tweet count. Note that this data is highly skewed into two distinct clusters.

al., 2010; Lee et al., 2011; Aiello et al., 2013). In the cases of non-localized trending topics, the hypothesis is largely supported, even with initial results as discussed earlier, when "initial" is defined as $log(n)/log(N) > 0.5$. Interestingly, and rather ensuring, these results are robust regardless of particular geocoder employed, which has substantive impacts as identified by Graham et al..

There are many potential biases present, which must temper any acceptance of this hypothesis. First, the trending topics that were collected, while from a total population, are biased to the time in which they were sourced – this is evident in the fact that there are "march madness" topics present within the sample. Attempting a similar study with either a more broad selection of trending topics or a selection of trending topics in a different time of year may reveal different results. Second, users who have geocodable profiles may be distinct from other users in a systematic fashion that could possibly be related to the underlying assumption of the hypothesis surrounding geographically versus non-geographically homophilous communities. Third, the analysis is based on an incomplete dataset owing to complications with data collection – with the full set, results may differ from what has been found, but the fact that statistical tests have shown to be relatively consistent in their findings suggests that probability is unlikely. Third, and particularly in the case of "initial" being defined as the first six hours of a trending topic, the incompleteness of the data

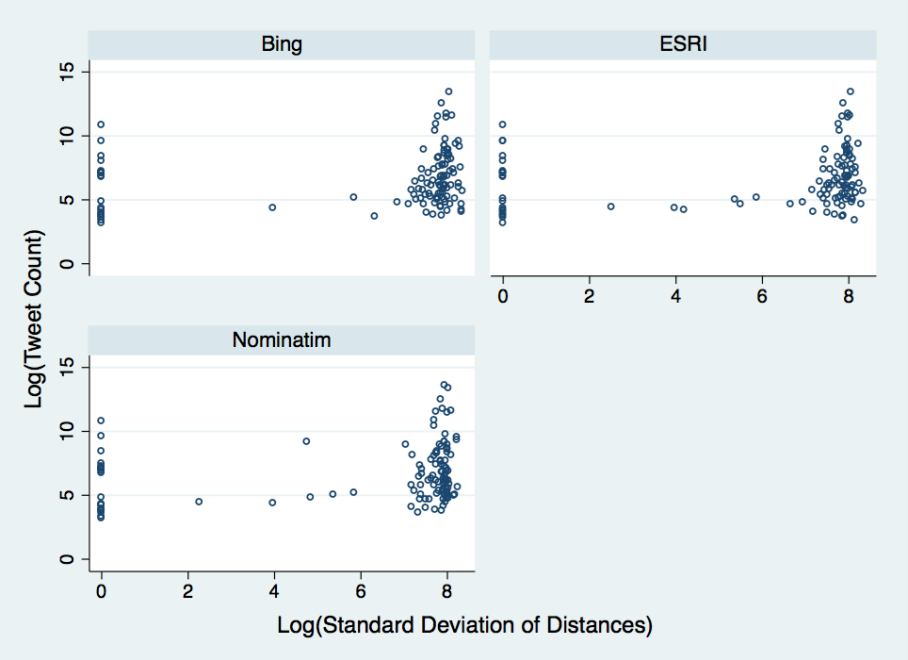

Figure 8: Scatterplot of log standard deviation of distances for all edges geocoded per topic by the log of tweet count. Note that this data is highly skewed into two distinct clusters.

collection in processing that definition of initial is a small dataset – further analysis may show that this definition is in fact significant, but such insignificant results at this stage lessen the probability that this may be the case.

Any study of data in this nature will likely carry heavy biases – the job of the researcher is to hypothesize and operationalize studies as robust as possible to these issues, and attempt to find results so significant that biases do not lessen the arguments laid forth. It is the hope that this result, along with the lengthy consideration of each part of the design of this study, as well as the substantively consistent results with the exception of localized trends, helps to mitigate these biases and show that in fact there is a real and useful relationship between topography and the popularity of a particular topic through the avenue of homophily and the particular ways in which the principle is found in online social networks.

N Log(Standard Deviation)				
Geocoder	$< k >> 1$	$log(n)/log(N) > 0.5$	$t > 6$	$N > 1000$
Bing	33	89	24	105
ESRI	30	86	24	99
Nominatim (OSM)	26	70	20	89
Spearman's ρ				
Geocoder	$< k >> 1$	$log(n)/log(N) > 0.5$	$t > 6$	$N > 1000$
Bing	0.0884	0.1002	0.2009	0.1144
ESRI	0.0227	**0.2530**	0.0191	**0.2297**
Nominatim (OSM)	-0.0315	0.1615	**0.5083**	0.1879

N Cross City Interactions				
Geocoder	$< k >> 1$	$log(n)/log(N) > 0.5$	$t > 6$	$N > 1000$
Bing	33	89	24	105
ESRI	30	86	24	99
Nominatim (OSM)	26	70	20	89
Spearman's ρ				
Geocoder	$< k >> 1$	$log(n)/log(N) > 0.5$	$t > 6$	$N > 1000$
Bing	**0.5488**	**0.6461**	0.0499	**0.5348**
ESRI	**0.4251**	**0.2901**	0.0984	**0.2756**
Nominatim (OSM)	**0.4466**	**0.3042**	-0.0231	**0.3200**

N City Diversity Interactions				
Geocoder	$< k >> 1$	$log(n)/log(N) > 0.5$	$t > 6$	$N > 1000$
Bing	33	89	24	105
ESRI	30	86	24	99
Nominatim (OSM)	26	70	20	89
Spearman's ρ				
Geocoder	$< k >> 1$	$log(n)/log(N) > 0.5$	$t > 6$	$N > 1000$
Bing	-0.1136	0.0619	-0.0245	0.0966
ESRI	-0.0847	-0.1208	-0.0554	-0.0584
Nominatim (OSM)	-0.0296	-0.1019	-0.3815	-0.1000

References

Aiello, L. M., Petkos, G., Martin, C., Corney, D., Papadopoulos, S., Skraba, R., ... Jaimes, A. (2013). Sensing trending topics in twitter. *Multimedia, IEEE Transactions on*, *15*(6), 1268–1282.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, *74*(1), 47.

Aragón, P., Kappler, K. E., Kaltenbrunner, A., Laniado, D., & Volkovich, Y. (2013). Communication dynamics in twitter during political campaigns: The case of the 2011 Spanish national election. *Policy & Internet*, *5*(2), 183–206.

Ardon, S., Bagchi, A., Mahanti, A., Ruhela, A., Seth, A., Tripathy, R. M., & Triukose, S. (2013). Spatio-temporal and events based analysis of topic popularity in twitter. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management* (pp. 219–228).

Asur, S., Huberman, B. A., Szabo, G., & Wang, C. (2011). Trends in social media: Persistence and decay. *Available at SSRN 1755748*.

Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 65–74).

Baños, R. A., Borge-Holthoefer, J., & Moreno, Y. (2013). The role of hidden influentials in the diffusion of online information cascades. *EPJ Data Science*, *2*(1), 1–16.

Benevenuto, F., Magno, G., Rodrigues, T., & Almeida, V. (2010). Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)* (Vol. 6, p. 12).

boyd, d., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662–679.

boyd, d., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on* (pp. 1–10).

Cairncross, F. (2001). *The death of distance: How the communications revolution is changing our lives*. Harvard Business Press.

Cairncross, F. (2002). The death of distance. *RSA Journal*, 40–42.

Cha, M., Benevenuto, F., Ahn, Y.-Y., & Gummadi, K. P. (2012). Delayed information cascades in Flickr: Measurement, analysis, and modeling. *Computer Networks*, *56*(3), 1066–1076.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM*, *10*(10-17), 30.

Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., & Leskovec, J. (2014). Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web* (pp. 925–936).

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political polarization on twitter. In *ICWSM.*

Erdős, P., & Rényi, A. (1959). On random graphs. I. *Publ. Math. Debrecen*, *6*, 290–297.

Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z., & Kellerer, W. (2010). Outtweeting the twitterers-predicting information cascades in microblogs. In *Proceedings of the 3rd conference on Online social networks* (pp. 3–3).

Gayo-Avello, E. M. M. S. H. S., Panagiotis Takis Metaxas, Peter Gloor, D., Castillo, C., Mendoza, M., & Poblete, B. (2013). Predicting information credibility in time-sensitive social media. *Internet Research*, *23*(5), 560–588.

Goel, S., Watts, D. J., & Goldstein, D. G. (2012). The structure of online diffusion networks. In *Proceedings of the 13th ACM conference on electronic commerce* (pp. 623–638).

Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer*, *66*(4), 568–578.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (pp. 56–65).

Kamath, K. Y., Caverlee, J., Lee, K., & Cheng, Z. (2013). Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 667–678).

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web* (pp. 591–600).

Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., & Choudhary, A. (2011). Twitter trending topic classification. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on* (pp. 251–258).

Lerman, K., Ghosh, R., & Surachawala, T. (2012). Social contagion: An empirical study of information spread on digg and twitter follower graphs. *arXiv preprint arXiv:1202.3162*.

Lotan, G. (2011, October 12). *Data Reveals That "Occupying" Twitter Trending Topics is Harder Than it Looks!* Retrieved 2011-10-12, from `http://giladlotan.com/2011/10/data-reveals-that-occupying-twitter-trending-topics-is-harder-than-it-looks/`

Macskassy, S. A. (2012). On the Study of Social Interactions in Twitter. In *ICWSM.*

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of Sociology*, 415–444.

Nagar, S., Seth, A., & Joshi, A. (2012). Characterization of social media response to natural disasters. In *Proceedings of the 21st international conference companion on World Wide Web* (pp. 671–674).

Nahon, K., & Hemsley, J. (2013). *Going viral*. Polity.

Nikolov, S., & Shah, D. (2012). A nonparametric method for early detection of trending topics. In *Proceedings of the Interdisciplinary Workshop on Information and Decision in Social Networks (WIDS 2012). MIT.*

Rattanaritnont, G., Toyoda, M., & Kitsuregawa, M. (2012). Characterizing topic-specific hashtag cascade in twitter based on distributions of user influence. In *Web Technologies and Applications* (pp. 735–742). Springer.

Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. (2011). Socio-Spatial Properties of Online Location-Based Social Networks. *ICWSM*, *11*, 329–336.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social networks*, *34*(1), 73–81.

Teske, D. (2014). Geocoder accuracy ranking. In *Process Design for Natural Scientists* (pp. 161–174). Springer.

Twitter. (2010, December 8). *To Trend or Not to Trend...* Retrieved 2011-12-8, from `https://blog.twitter.com/2010/trend-or-not-trend`

Twitter. (2014). *What are Promoted Trends?* Retrieved from `https://support.twitter.com/articles/282142-what-are-promoted-trends`

Weng, L., Menczer, F., & Ahn, Y.-Y. (2013). Virality prediction and community structure in social

networks. *Scientific reports*, *3*.

Yardi, S., Romero, D., Schoenebeck, G., et al. (2009). Detecting spam in a twitter network. *First Monday*, *15*(1).

Zipf, G. K. (1949). Human behavior and the principle of least effort.

Zubiaga, A., Spina, D., Fresno, V., & Martínez, R. (2011). Classifying trending topics: a typology of conversation triggers on twitter. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 2461–2464).