Determining the Importance of Geography on Twitter

Candidate Number 791026

Thesis submitted in partial fulfilment of the requirement for the degree of MSc in Social Science of the Internet at the Oxford Internet Institute at the University of Oxford

August 1, 2012

Word Count: 9,769

Abstract

The effective annihilation of geography as an intrinsic consideration in the transfer of information is one of the most salient features of the modern communications landscape. That a modern information and communications technology such as the Internet has this potential capacity, though, does not necessarily mean that it is incorporated into its users activities – the core point of interest, then, is assessing the degree to which users transcend physical space when interacting online. While much theoretical work exists, empirical evidence typically suffers from problems of precision of geographic measurement. Twitters voluminous streaming data API allows for a large swath of tweets to be collected with ease, and has proven to be a useful data source for Internet research. By analyzing a novel data source, geolocated tweets, precise geographic measurements can be employed and assessed, and used to measure the salience of geography at much finer detail. This work aims to address problems of data precision in existing literature discussing the salience of geography in interactions on Twitter, and claims that geolocated data allows for a different angle from which to view the question. A sample of tweets from a large corpus of raw data are employed, and a model is established which tests the degree to which geographic, semantic, and social distances impact the frequency of interactions between users. Ultimately, while geography still matters, it's effect is minimal to a point where it is not substantively important when compared to semantic distance between users. The work contributes to the existing literature by adding an additional methodological path towards modeling of geographic importance, and addresses the most significant methodological concerns with online research in the process.

Introduction

The notion that a communications medium's transformative powers in altering the geographic contours of our lives is a pervasive narrative throughout the history of Information and Communications Technologies (ICTs) generally – its application to the Internet should be no surprise. What is new, however, is the articulated plasticity of this medium relative to its predecessors – while earlier ICTs were capable of exhibiting a wide range of use, the ease, scale, and speed at which the Internet can incorporate and appropriate a wider diversity of applications is central to its rise as a dominant mode of communication and information seeking.

Specifically, a perceived central role of ICTs are their apparent ability to collapse geographic distance by allowing users to maintain ties with physically distant others – by providing better methods of communication between people, the logic goes, considerations of geographic distance will weigh less for people, and geographic boundaries are able to be transcended. It is relatively uncontroversial to agree that personal geographic expansion due to reduced communication costs is at least theoretically possible. If true in practice, however, various corollaries fall from this, carrying profound implications – it is in the space that the assumptions driving Friedman's (2005), Cairncross' (1997), and other's work are found. But the assumption made, that this theoretical possibility is practically realized, is far from certain. As will be discussed, numerous attempts have been made to assess aspects of the practical contours of this theoretical removal of geography – it is in this literature that the current work is primarily ensconced. Broadly, this work seeks to add further definition to the variables involved in shaping the practical contours of an online user's personal network.

As in Mok, Wellman, and Basu (2007), the frequency of interaction, geographic distance during interaction, and underlying relationship between those who interact constitute the three primary dimensions of analyzing relationships between users. While there are theoretically innumerable factors contributing to this un-

derlying relationship dimension, it is argued that the most salient factors for a user is their semantic and social proximity to the users they interact with. In other words, while geography may play a significant role in the frequency of interactions between users, other factors, such as interest similarity (eg. sustained posting of similar content) or pre-existing social ties (eg. friend, kin, and work social networks) may help explain much of the variance seen.

The goal of this work, then, is in part to add nuances to the practical contours of the role of geography in online networks. Certainly, Scellato, Noulas, Lambiotte, and Mascolo (2011)'s work, amongst others, have done some of this before – the goal is not to simply re-assert old findings with identical methodologies. Specifically, with Twitter's nascent geo-coded API, an assessment of location-based usage of a popular social network service is possible to a novel degree of geographic certainty. The paper also aims to move beyond similar research conducted by Takhteyev, Gruzd, and Wellman (2012), and argues that far from structural elements underlying communication between two users, social and semantic distances between users are vital parts of the equation, and must be discussed in conjunction with geography. It similarly extends work by Scellato et al. (2011) to include this semantic dimension, as Twitter's status as a directed network and culture of link exchange promotes ties between actors with shared interests, which may in turn have explanatory power for connections made on the network.

In assessing this refinement of geographic contours, the work will begin with a literature review, including the various related fields that may lay claim to this question, and will take care in determining the most efficacious approaches utilized thus far. Then, the methods section will lay out an approach that stresses high internal validity over generalizability to the rest of Twitter or the Internet. Results from the study carried out will then be presented and subsequently discussed.

Definitions

A few terms have been used casually up to this point, but must be more specifically enumerated. Most basic is Twitter itself. While Twitter is a relatively known quantity, depending on the context of a study, it may play vastly different roles for the users that communicate and seek information through it. On the most basic level, it is simply a directed-link online social network (OSN). As a directed OSN, it has properties that do not necessarily compare to undirected-link OSNs, the most important being the ability to cultivate large followings – as a result, a major component of Twitter is the predominance of major accounts (Cha, Haddadi, Benevenuto, & Gummadi, 2010). Beyond the structural nature of this OSN, the messages sent must also be defined – in this work, the statuses posted (tweets) are restricted to mentions and retweets, both fundamentally conversational activities between users theoretically exposed to wider audiences (boyd, Golder, & Lotan, 2010). Notably, this study is focusing on the "user", which is defined as an account on Twitter. As Chu, Gianvecchio, Wang, and Jajodia (2010) note, "users" on Twitter fall into three types, "human", "cyborg", and "bot", where cyborg is defined as "either [a] bot-assisted human or human-assisted bot" . The distribution they find for humans, cyborgs, and bots is approximately a 5:4:1 ratio, which leaves considerable space for semi-automated and completely automated accounts to be present in this, and indeed any, Twitter dataset (2010). The term "user" is then preferred to encapsulate these actors, and generally considered to be human or human-based actors.

Within this study, geographic, social, and semantic relationships are considered in relation to the frequency of contact. In Mok et al. (2007)'s conception, this is akin to framing social and semantic relationships as observable underlying relationships between users. While there are certainly many other forms of relationships, such as the macro-structural ones considered by Takhteyev et al. (2012), this study seeks to define the geographic contours as they occur within

Twitter's site – inviting such external variables is a step away from the intent of this research. Of these three variables – geographic, social, semantic – geographic is relatively straightforward as a direct distance between actors (rather than the more complex measurements as offered by considering routes, modes of transit, and temporal constraints as in work by (Thrift, 1977), as more nuanced measurements require unobservable knowledge about the actors themselves). Social relationships are defined in the context of a user's ego-network, or the connections that users, their friends, and their followers maintain with one another. Specifically, social relationships are defined by the number of mutual friends and followers that two interacting users share, and is a relative metric across varying interactions. Semantic relationships are similarly defined as a dimension of the underlying relationship between the two users. In this case, the historical content each user tweeted is compared for semantic overlap – the higher degree to which users talk about the same concepts, places, people, and the like, the more semantically similar they are – in other terms, this is also a relative metric for determining the degree to which their expressed interests overlap.

Literature

This paper is in many respects a response to previous attempts to understand the practical geographic contours of communication online, specifically through Twitter. As such, the literature considered is as much the foundational assumptions and general theoretical work surrounding the question of communication patterns over geographic distance online as it is the specific attempts to explicate that relationship.

Intervening Opportunities

The single most predominant academic debate being engaged with is the fundamentally sociologically-driven effort to understand the role that the Internet specifically, and communications technologies more generally, plays in eliminating geographic distance from communication. Stouffer (1940)'s concept of "intervening opportunities", which is the notion that "the number of persons going a given distance is directly proportional to the number of opportunities at that distance and inversely proportional to the number of intervening opportunities" (846) is a particularly salient point of entry. Simply put, regardless of any goal, the further a person travels to realize their goal, the higher the cost to realize that goal – distance is an inevitable cost in any benefit one derives from activity in physical space. While Stouffer (1940)'s work focuses on the migration patterns of families in Cleveland, the general theory is said to be possibly applicable in many cases.

The impact of ICTs on geographic considerations for connections between people who communicate is then a derivative discussion – when the cost of communication relative to the distance between actors decreases, while still maintaining some level of richness, do actors actually scale out beyond their immediate geography? That is, does a person, when capable of talking to anyone at great distances, actually talk to them? Of course, Stouffer's model actually argues that the increased travel must correspond to increased benefit – it may be that the actors of most use tend to be geographically close. Still, probabilistically, geographic spread should increase if the theory held.

Much work has been done surrounding pre and post-Internet geographic shifts for communicating actors, and work has tended to temper the more exceptionalist arguments that the Internet may uniquely change this relationship in a way no previous ICT ever could have. To be fair, few people operate on the absolute fringes in the contemporary debate – it is as unfounded to proclaim the Internet as the catalyst for the "end of geography" as it was to proclaim the end of

the cold war as the catalyst for the "end of history" (Bates, 1996; Fukuyama, 1989). The debate is instead primarily focused on the practical extent to which traditional geographically-bounded relationships are augmented and new relationships formed through the new forms of communication offered in theory by the Internet, and as such, should be seen as a specific derivative of Stouffer (1940)'s "intervening opportunities" theory.

Geography

Determining the importance of geography on Twitter, then, is the central goal. Specifically, the unit of analysis in a question driven by Stouffer (1940)'s theory is the individual – as such, the focus of literature is on geographic analysis as it results to the individual actor. Geography's importance is paramount – Daraganova et al. (2012) are careful to point out geography's resilience in the face of many iterations of technologies that increase geographic reach. Of existing literature, however, one of the most theoretically useful frameworks can be found in Miller McPherson and Cook (2001)'s work on homophily. Broadly, the concept of homophily is that any given actor expresses a number of distinctive attributes – geographic position, age, race, gender, and so forth. In a social network, these attributes serve a role as anchoring similarities between actors, who in the aggregate organize around these similar traits. Miller McPherson and Cook (2001)'s work is an exhaustive review and defense of this framework, and argues that above any attribute expressed by actors, the most "basic source of homophily is space" (429), or geography.

One of the more instructive methodological offerings in this niche is provided by Mok et al. (2007). In their work, the frequency of contact between actors is held as an independent variable, and the geographic euclidean distance and underlying relationship (operationalized as combinations of kinship status and self-reported intimacy levels) are used as dependent variables. This method is a

fairly direct approach to assessing the impacts of multiple variables on frequencies of contact, and was replicated in a related work by others. In Takhteyev et al. (2012)'s analysis of Twitter data, kin/intimacy relational variables are substituted with more structural relationships between the regions users were located in.

Beyond general approaches for measuring the salience of geography, much work has been done in assessing the empirical relationship on various online networks. Notably, Liben-Nowell, Novak, Kumar, Raghavan, and Tomkins (2005)'s work on LiveJournal is instructive in setting possible expectations for findings. Briefly, LiveJournal is a blogging platform that, like Twitter, features directed ties between users, but unlike Twitter, engenders a much more tight-knit community due to the long-form posting that occurs on the platform. In their work, they observed that the probability of links between users decreased as the unit of distance (in this case, kilometers) between them increased to a point where the probability leveled out to a relatively constant and low value at distances roughly exceeding 1,000km (Liben-Nowell et al., 2005, 11625).

Social Ties

The single most instructive work on social ties guiding this research is Granovetter (1973)'s work on the strength of weak ties, particularly the forbidden triad. In his assessment, the reader is instructed to consider three actors, A, B, and C. If A and B are connected, and A and C are connected, then it is likely that B and C are connected as well. This transitivity is argued as being a fundamental component of human relationships – simply put, it is hard to maintain affiliations with two actors who actively reject a direct relationship between themselves, and more often, there is a relationship when the others are strongly connected. When explicit, this complete three node graph denotes a fundamental component of social networks, the triad. By extension, Granovetter posits that the presence of a "forbidden" triad are cases where the connections are not strong – through his

careful analysis of the fundamental properties of network structure, he is able to find that, by definition, a bridge tie will not be part of a triad, and thus, must be a weak tie. Granovetter expands this model, and argues that these weak ties are the primary structural nuance responsible for the transitivity of information as seen in Milgram (1967)'s "Small World Problem". As one of the most cited works in the social sciences, Granovetter (1973)'s work is of paramount importance for understanding how and why humans communicate with others.

Granovetter's salient contribution for much of the research done on varying aspects of Twitter focuses on these weak ties, and this research aims to build on this work. Notably, Golder and Yardi (2010)'s work on transitivity and mutuality of ties on Twitter is the most direct engagement of Granovetter's work with weak ties as it pertains on Twitter. The authors collected lists of second-order followers (or their followers' followers'), and asked users to assess the degree to which they would friend the users. As an additional validity check, they added a follow button that would allow for the user to actually begin following the user in question. Without the explicit information, the authors posited, the users would not be as willing to connect to these users. Remarkably, Golder and Yardi (2010)'s work provides no control where users are shown their alter tie – nonetheless, they conclude that without tie information, users are reticent to connect. A more directly translatable work, however, is Lussier and Chawla (2011)'s analysis of social ties based off of the KAIST dataset used in Kwak, Lee, Park, and Moon (2010). In this work, the authors examine mentions, retweets, and link exchange against a backdrop of social ties between the users. In their analysis, they find empirical confirmation of Granovetter (1973)'s theory – users with low clustering coefficients, or users that are embedded in any single "community" are more likely to be retweeted. While there are certainly conflating variables involved, notably the number of followers a user has (if a user has many followers, it intuitively follows that the following users will be spanning multiple communities), the result

points in the theorized direction. Another interesting property of social ties, elucidated in Scellato, Musolesi, Mascolo, and Latora (2010), is the robustness of triads despite vast geographic differences – in their comparative study of multiple social network sites, the authors found that "geographic distance influences only the geographic properties of the triangles, not their likelihood of appearance".

In the scope of this research, transitivity will have a related role – in assessing the union of egonets of users, it will be assumed that implicit triads – cases where users are mutually tied to other actors in the network – will increase the possibility of communication between users. While in practice the presence of direct connections between users who communicate in this dataset will likely be significant, these additional bridging cues may provide useful insight into the relative connectedness of users.

Semantics

Interestingly, Miller McPherson and Cook (2001)'s work also points towards an interest-based approach to assessing homophily, albeit at a much lower explanatory power in their theory. In their work, their literature review identifies that across many works looking at social networks, primary catalysts for homophily, such as geographic position, age, sex, gender, race, and ethnicity, manifest in derivative ways: "occupation, network position, behaviors, and intrapersonal values also show considerable homophily, but they seem to be more specific to certain types of networks and/or derived from the basic facts of sociodemographic homophily" (Miller McPherson & Cook, 2001, 429). In their own analysis of the factors influencing homophily, the authors point towards "[s]chool, work, and voluntary organizational foci" as sources of homophily, in addition to other cohort-focused organizational schemes, such as employees at the same level of employment in a workplace or ties induced through marriages, where spouses serve as bridges to other groups (Miller McPherson & Cook, 2001, 435). In practice, controlling

for possible collinearities of these phenomenons, i.e. assigning relative weight to each of these types of sources of homophily, is dangerous – by their nature, these relationships are self re-inforcing. What can be done instead, is to look at the results of homophily – by assuming homophily plays a role in social tie formation, it may be posited that the similarity of people may manifest itself in the nature of a person's content online. The result, then, is that users who may be more geographically diverse may remain tied to one another either explicitly through the network or implicitly through the similarity of their content – if users play profoundly similar roles at geographically distant locales, and find themselves expressing similar age, race, gender, ethnicity, class, and organizational attributes, one may expect that geography may be more easily transcended. For this reason, a semantic value, as a proxy for measuring the interest and content similarity of users, is used to determine whether or not content-based homophily plays a role in augmenting simple geographic distance contours.

Unfortunately, little work has been done in assessing the utility of semantic similarities between users – this is likely a result of the richness of other available data, particularly the friend/followership lists which would provide a much better proxy in most cases. In this case, however, the research aims, in part, to disentangle the relationship between the semantic and network similarity between users, in addition to geographic similarity, rather than use them as a basis for other assumptions. Of the work that has been done, most focuses on profiling users for purposes of user classification through semantic analysis, or for inferring latent variables such as age, gender, and educational status (Tinati, Carr, & Tarrant, 2012; Rao, Yarowsky, Shreevats, & Gupta, 2010; Hauff & Houben, 2011). The current work, however, is interested not in latent variables or generalized profiling, but instead the utility of semantic similarity as a component for explaining increased frequency of communication between actors. One work that approaches tangentially related aims is Pennacchiotti and Popescu (2011)'s

work on classifying user types. While this work is again a profiling-based work, in a significant section of their work, they aim to profile users as either Democrats or Republicans. In this effort, they seed a "gold" standard of Democrats and Republicans by using WeFollow.com and Twellow.com, services that allow users to generate a Folksonomy around users on Twitter. In other words, these services allow users to tag other users, including themselves, as tweeting about or on multiple themes, concepts, or industries (eg a user can be tagged as someone who regularly tweets about US Politics, Democrats, and Washington, DC restaurants). From this, the authors collected egonets of these seed users, and used a battery of machine learning algorithms to classify alters as either Democract or Republican. In their work, though the focus was defining the efficacy of different algorithms, the "gold" standard seed data seemed to be particularly relevant to the approach that could be taken in this research, though in practice, the data from these services is minimal to the point that they do not achieve the goals that are desired by adding a semantic dimension to the model. As such, an alternative approach that approximates such services, the number of mutually-employed hashtags, is discussed at further length below and ultimately employed.

Previous Work

A few works focus specifically on assessing the geographic contours of Twitter communication, though in sum their contributions still necessitate the current research. Takhteyev et al. (2012)'s work on this topic is most similar – as mentioned above, it largely borrows Mok et al. (2007)'s methodology of frequency of contact measured against geographic proximity and underlying relationships. In this work, the authors investigate structural variables in place of an actor's personal relationship with others – specifically, the authors use the presence of shared national borders, direct flight connections, and language overlap as proxies for estimating the proximity between actors. The paper, while interesting, suffers

from a few flaws – most notably, the unit of analysis is ambiguous. Though the authors are intent on describing individuals, they collapse individuals into regions, and conduct their analysis on a region as their unit of analysis – flights connect regions, not individuals. Additionally, the presence of language similarities, while true in the aggregate, need not be true for the individual, and while the presence of a border may indicate shared ties based on larger regional cohesion (eg Canada and the United States), it again does not necessarily hold for the individual users studied. While employing structural variables is theoretically compelling, it suffers from possibly being an ecological fallacy, where users that were considered may have little connection to these underlying regional variables, and other variables may in fact explain the phenomenon much better (Babbie, 2010). Still, the paper stands as an interesting attempt to assess Twitter's geographic contours – ultimately, the authors find that the presence of direct flights between regions is the largest determinant beyond geography in explaining the frequency of communications between actors.

Scellato et al. (2010)'s work focuses instead on explicating the geographic distributions between communications on different services. Their core argument is that different OSNs have vastly different aims, audiences, and activities associated with them – beyond that, some OSNs are directional, while others require mutual friendship. In their assessment, they find that different OSNs, that is, different contexts for users to engage in communication with one another, exhibit different properties – included in this study is Twitter, which shows the widest geographic spread. While the statistics are not rigorously conducted, the authors cite that only 5% of friend/followerships are between users located within 100km of one another – the average distance is reported as 5,117km. While the work is relevant, their sole focus on friend/followership between users makes it difficult to discern which friends are actually communicating, which, again, is the interaction being studied.

Finally Kulshrestha, Kooti, Nikravesh, and Gummadi (2012)'s work provides a hybridized study on geographic contours of Twitter users – the authors consider structural factors as they relate to the friend/followership relationships between geocoded actors. In their analysis, they focus on a concept of "importing" and "exporting" actors, positing that some nations play roles as producers and others as consumers of content on the platform. In this effort, they similarly identify structural relationships, such as the presence of linguistic similarity, geographic closeness between nations, and the presence of shared national borders, all in conjunction with environmental factors such as national metrics (eg. the Human Development Index) as a variable estimating the digital divide. While an interesting approach, again, the assessment possibly suffers from an ecological fallacy, as it conflates variables of varying units of analysis in the study – while the Human Development Index may serve as a proxy for the severity of the digital divide between countries, it cannot provide that on an actor level (Babbie, 2010). Additionally, the concept of export/import and in their ultimate assessment, tweet surpluses and deficits, is at best tangential to the underlying intent of this work, which purports to address the salience of geography.

Methods

Provocations for Big Data

(boyd & Crawford, 2012)'s recent work, outlining six major "provocations" for such studies, is an excellent outline for the primary methodological concerns for any paper dealing with large-scale data mining of online social network data. In describing the methodology for this project, (boyd & Crawford, 2012)'s work will be used as a general roadmap for the types of issues the work will attempt to address and mitigate.

Automating Research Changes the Definition of Knowledge In their work, (boyd & Crawford, 2012) argue that the "computational turn" seen in social sciences, and particularly Internet studies, is more than a simple change in methodology, but instead changes the fundamentals of knowledge production. They cite Wired's Editor-in-Chief Chris Anderson vision of "letting the numbers" speak for themselves without concern for theory as an example of the danger of this computational turn (Anderson, 2008). True, it would not be defensible for a researcher to approach problems without theory in hand – the most successful applications of big data, however, maintain the position of theory. Wu, Hofman, Mason, and Watts (2011)'s attempted confirmation of Katz and Lazarsfeld (1955)'s two-step flow theory is a perfect example of theoretically-informed large scale data mining on Twitter. Wu et al. (2011)'s work on "intermediaries" on Twitter are a direct theoretically-driven case study, and effectively avoids much of the concern that boyd and Crawford have of entirely novel methods of knowledge production. Similarly, the question at hand aims to begin solidly from theoretically-driven interest, and implements a study through big data.

Claims to Objectivity and Accuracy are Misleading Here, boyd and Crawford argue that because a study is conducted at a large scale and deals with quantifiable data does not mean that it is necessarily true. In this effort, boyd and Crawford are not pointing out inherent issues with this type of research per se, but showing the perhaps willful ignorance of some researchers to respect the role of interpretation. To boyd and Crawford, even selecting salient variables is problematic as it is a form of interpretation – this methodological concern is fundamentally about a perceived tendency to over-estimate the abilities of big data research. Specifically, some form of big data exceptionalism provides auspicious confidence in its representativeness and generalizability. As boyd and Crawford argue, "[b]ig [d]ata is at its most effective when researchers take account of the

complex methodological processes that underlie the analysis of social data", that is, when researchers take care to contextualize, parameterize, and define online data as it pertains to a social theory being investigated (boyd & Crawford, 2012, 6). This idea has been a part of literature for quite some time - Scott specifically warned of "the mathematical tail wagging the sociological dog" in network theory research (Scott, 1988, 112). This problem cannot be reduced, however - it can only be mitigated through careful treatment and framing of conclusions, contingent upon what is found. As such, any findings stemming from this research are parameterized to the Twitter landscape - indeed, only those interactions which bear the same attributes as those included in the final model.

Bigger Data are Not Always Better Data Here, boyd and Crawford caution that more is not inherently better, and that regardless of size, particularly in the scope of Twitter research, "when researchers approach a dataset, they need to understand – and publicly account for – not only the limits of the dataset, but also the limits of which questions they can ask of a dataset and what interpretations are appropriate" (boyd & Crawford, 2012, 7). Fundamentally, the objection at hand is that large scale analysis is not necessarily as internally or externally valid as a small dataset – size can only do so much. While a large sample size permits increasingly specific questions about distinct subsets of users (since the number of records or respondents matching increasingly specific conditions necessarily decreases), the goal of this study is to observe a general trend across users. While a more general picture of the entirety of the 7,925,128 tweets in the raw catalog may give a slightly more precise sense of the relationship, the trade-off is a significant decrease in the complexity of questions that can be practically asked. For this reason, only a small sample is queried, which serves as an estimate for the population.

Not All Data are Equivalent boyd and Crawford carefully distinguish between behavioral and articulated networks as the types of networks that can be measured on social media systems as distinct from personal networks, the types of networks that are more so in the domain of classical sociological work (Fischer, 1982). This point is important, but does not particularly have any methodological concern in the research being considered – this research is solely interested in the impact of geography on the behavioral and articulated networks, which, while they may have an overlap with the personal network, is beyond the scope of asking very broad and general questions of the relationship between information flow online and user geography. In other words, this research is not trying to read further into sociological notions of geographic determinism - the main concern here is relationship itself, not the causes – the expressed, rather than the personal, network is the focus of the research, and as such the work only focuses on questions that can be directly answered via the data collected. In a more general sense, the variables used in this study are directly observable, rather than latent, data concerning users and their tweets. In particular, probabilistic models of semantic distance (by employing methods such as Deerwester, Dumais, Furnas, Landauer, and Harshman (1990)'s latent semantic analysis) are of much practical use – the overarching goal in this work, however, is to observe behavior, not to model it.

Just Because it is Accessible Doesnt Make it Ethical Reminiscent of Nissenbaum's notion of contextual integrity (2004), boyd and Crawford cite the Facebook network ties research conducted by Lewis, K. and Kaufman, J. and Gonzalez, M. and Wimmer, A., and Christakis, N. (2008), and the failure to adequately privatize the data that was collected. Because of the unique network structure of the users considered, it was possible to de-anonymize the dataset, and, in effect, expose private data in a way that could not have been anticipated

by the subjects being researched. Ethics can and should play a role – after all, boyd (2007)'s work on MySpace raises the point that "invisible audiences" change notions of distinct separations of public and private life online – just because its publicly accessible does not mean it is meant to be shown to the whole world. For this reason, this proposed work makes a concerted effort to address this ethical issue by categorically not disclosing personally identifying information – all results considered are considered in the aggregate. This work goes beyond previous research on Twitter, though, in addressing the possibility of temporary tweets. While a tweet may be fully public, it can also be deleted at will – this creates a nebulous ethical problem, where research could be conducted on data that has since been deleted or privatized. As such, all tweets used in this research were reviewed several months later to ensure that the tweets persist in the public user's timeline. In general, current work in the field seems to allow for a high-level review of public data gleaned from Twitter while remaining ethical, as is seen in numerous works (Wu et al., 2011; Kwak et al., 2010; Cha et al., 2010; Chu et al., 2010).

Limited Access to Big Data Creates New Digital Divides For the authors, the phenomenon of data "insiders" and "outsiders" – those who do and do not have computational backgrounds, maintain expensive contracts with data providers, and have computational resources to throw at questions – creates a divide between researchers that threatens unbiased intellectual output. As many data providers are proprietary profit-oriented outfits, such as Twitter and Facebook, there exists a clear incentive to favor research that favors their ultimate valuation. This work can address this problem however, by using the lowest-level of access to Twitters data store. By doing so, the data collected will be free of any higher-level restrictions (though it will not be as complete as a result), and will be collected in a way that presents the fewest barriers for future researchers to repli-

cate the work and confirm or refute the results produced in a methodologically comparable fashion. Specifically, this dataset will be based off of the "spritzer" method of Twitter streaming data access level, a publicly available resource for any potential future researchers.

Geocoding Data

Also central to this paper is a fundamental challenge to previous works surrounding geographic analysis of Twitter data. All previous works exploring geography on Twitter share a common and hitherto unavoidable methodological flaw, in that they rely on the accuracy of self-reported locations. On Twitter, a self-reported location is an open-ended text field that is offered to a user. When a user edits this field, they are prompted with the question of "Where in the world are you?" (Hecht, Hong, Suh, & Chi, 2011). This location data suffers from many methodological problems, and as such, much work done on locational questions are written under improperly conditioned assumptions.

The only work aimed at explicating the degree to which self-reported locations are incorrect is Hale, Gaffney, and Graham (2012), which focused on four metropolitan locales "characterized by interesting geographic, linguistic, and cultural differences", San Diego, Montreal, Cairo, and Tokyo. The data collected for this study was only from users who tweeted using device locations, so that it was possible to directly measure the difference between algorithmic coding self-reported location and what was considered to be a precise location. In their analysis, even in the best cases, the median distance for one algorithm, Yahoo's PlaceFinder, was 5.4 miles, while the worst-performing case was seen in PlaceFinder's analysis of Tokyo, exhibiting a median of 16.3 miles. In other words, when the self-reported location information provided by users in Tokyo was fed into PlaceFinder's location estimation algorithm, only 50% of the users locations differed less than 16.3 miles from the exact locations delivered by the

GPS coding of their location at the time they posted.

While the numbers arrived at by Hale et al. (2012) are contingent upon the regions studied, they provide a useful general sense of the scale of error one may expect in estimating true location from self-reported locations. As any act of communication on Twitter happens between two users, one may expect the upper bounds of error to then be somewhere along the order of a few dozen miles, which would correspond to an error of expecting users to be located in the same neighborhood when their true locations place them in neighboring cities instead. Of additional importance, not addressed by any current research, is a phenomenon very similar to a known issue on the web, "link-rot". Link-rot is the phenomenon where web links are prone to decay as the external resources they point to are moved and removed, subsequently eliminating any usefulness of pointing to the resource (Bar-Yossef, Broder, Kumar, & Tomkins, 2004). Similarly, while a user may self-report living in "Boston, MA, USA", it's entirely plausible that the user may move to some new locale, not update their Twitter profile to reflect this new change, and subsequently the location will be accurately geocoded, but inaccurate as to pointing to the real location of the user. No related work that utilizes self-reported locations attempt to answer this specific question, although it is tangentially addressed via Hale et al. (2012)'s work. What is clear, though, is that assuming it to be a negligible error, especially given the tweet-level accuracy device-driven data provides, is not methodologically safe.

Methodology

With these discussions in mind, it is possible to now discuss the particular operationalization of this study. Over the course of several months of data collection, tweets were collected via Twitter's streaming "spritzer" level of access, according to random stream of up to 1% of all data posted on the service. The data was specifically collected through the service's "filter" method, which allows for a

bounding box, or a set of two latitude and longitude pairs to be drawn on a map – in this case, the "bounding box" drawn was "-180.0,-90.0,180.0,90.0", which corresponds to a bounding box that includes the entirety of the planet. As a result, only tweets encoded with device-driven locations are included in this set, which, in its totality, comprises 195,923,943 tweets. As is stated in Hale et al. (2012), the portion of tweets that are geocoded lies somewhere below the 1% range (the authors specifically state it as 0.7% in their trial estimate), though that is an aggregate figure – at some points, it is higher, and at some points lower. As such, a minimal amount of data loss was seen in moments where the service was rate limited due to more than 1% of tweets being geocoded. As in Takhteyev et al. (2012)'s work, traffic increases peaked around the American east coast's afternoon schedule. As a result, if the data is truly random (no independent verification of Twitter's assertion is available), data will indeed be slightly over-represented in regions that tend to not tweet during the American east coast time-zone's afternoon, with likely spillover into the eastern seaboard's adjacent timezones.

From the entire corpus of data collected, 21 consecutive days (April 9th through April 29th, inclusive) of data were selected as the point of focus, as small errors occurred intermittently resulting in non rate-limit-based data loss several times - this selection of 21 days is free of any of these class of errors, and is of a length long enough to generally define these users as "currently active" users of the platform, without any sampling-based time-zone or day-to-day biases that may be seen in smaller sets. From this set, a further restriction was placed on the data – only tweets that were conversational in nature are considered. That is, only tweets that either retweet or mention at least one other user are included. Beyond this restriction, a further restriction was placed on the dataset – the users being retweeted or mentioned had to appear in the corpus as well in order to be included, so that their location could also be accurately mapped. Substantively, these restrictions ultimately mean that the users being examined are active geo-

locating Twitter users who interact with other active geo-locating Twitter users. While these conditions effectively bias the study to a degree that the findings are not generalizable to the entirety of Twitter, the trade-off of high internal validity of locations is specifically required in assessing the question at hand. Over the course of these 21 days, a total of 7,925,128 tweets matched these conditions. While a study could technically review this full set, for various practical considerations of scale , only a sample of this data is considered so as to perform more in-depth analysis while still remaining representational of the entire set.

An important decision point is whether or not this sample should be based on users or tweets. In many ways, both the user and their tweets could be considered as the individual unit of analysis – is the concern the actor or the interaction? Practically speaking, sampling based on user and sampling based on tweet generates vastly different data – while a user sample provides the full richness of distances users interact at (in that the sample would then include multiple communication acts between the multiple people they engage with), it is not necessarily representational to the larger dataset considered as directly. For this reason, while the user-based sample may more directly respond to questions of actors, the unit of analysis considered is the tweet and response pair, which drives many of these other decisions leading up to sample selection.

Geographic Distance "As the crow flies", while simple to calculate for two pairs of locations, is problematic when it comes to real interpretation. Beyond the physical Euclidean distance, work such as Thrift (1977) show that "actual" distance is much more complex. In the case of Takhteyev et al. (2012), categories of distances were used to describe the general difficulty in physical transit as a proxy for "actual" distance. In this research, distance is dealt with as the log of

[0]Specifically, the current API access available to researchers as it pertains to collecting friend and follower data, as well as historical tweets from users, makes it prohibitively temporally and computationally expensive to collect and store this data at full scale.

Euclidean distance as conducted in Liben-Nowell et al. (2005).

Social Distance As discussed earlier, Granovetter (1973)'s work on social networks is instructive in assessing how people communicate with one another for varied purposes. Additionally, Twitter's API-accessible friend and follower network data allows for researchers to directly assess who is connected to whom. By using Twitter's API, Granovetter's work can be used as a starting point for assessing the strength of ties between actors who communicate on the platform by looking at the larger social context in which these communication acts occur. While social ties are highly contextually bound, the relative strength of social ties across communication acts can be used as an instructive data point. In more operationalized terms, this work employs the number of shared ties in two user's friend and follower networks as a proxy for assessing the presence of "implicit" triads, and uses this count as the data point at which to assess social distance. The unit, then, is the number of implicit triads, or number of shared friends, between the tweeting user (the sample dataset) and the user to which they are responding.

Semantic Distance Semantic distance is by far the more difficult metric to quantify in as concrete terms as geographic and social distance. Whereas the variables for both of these are articulated by the user's metadata, user-articulated semantic metadata is sparse (such as the use of hashtags, urls, and user names), and introducing a non-articulated variable (by employing some semantic analysis technique such as Deerwester et al. (1990)'s Latent Semantic Analysis), while possibly helpful, runs a risk of deviating from the user's observed behavior in a work which seeks to describe and explore this behavior, rather than model it. For this reason, semantic distance is operationalized through user's uses of hashtags in their Tweets. Specifically, the union of hashtags between users – the user who posted content, and the user who replied to that content – is used as a count

variable indicating increased semantic closeness, in that hashtags are popularly used as assertions of a tweet's categorization or underlying meaning (boyd et al., 2010).

Sampling and Operationalization As a result of the above considerations, a random sample of 10,000 tweets were selected from the database, which was a large enough sample to ensure that in the event of missing, omitted, or deleted tweets, the sample would still be large enough to conduct a meaningful analysis. Each tweet in the sample was a response to another geolocating user's tweet – as such, their information was also collected as the "responded" tweet to which the sampled tweet was "responding" – individually, these are referred to interactions. For all 10,000 tweets, a process was run to ensure both that this specific tweet had not been deleted, and that both users had not since privatized their accounts, which yielded a subsequent set of 8,701 "interactions" between users. From here, geographic distance was simply calculated using a basic trigonometric function that returns the distance in miles. Social distance was calculated by determining the count of the union of mutual ties for the "responded" and "responding" users. For both users, the list of mutual ties to other users (situations in which the user follows and is followed by another user, as Twitter is a directed network) were compared, and the number of cases where mutual ties were extended to both "responded" and "responding" users, each being indications of implicit triads, was chosen as the measurement for social distance. Semantic distance was similarly assessed – the union of employed hashtags for a period of 28 days prior to the sampled interaction for both "responded" and "responding" users was chosen as an indicator of semantic distance. The dependent variable, total interactions, is the number of times that, in the course of that 28 day period, the users sent and received replies from one another.

Results

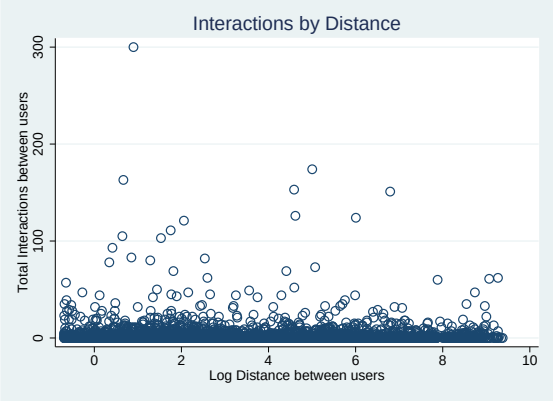
Formally, this research aims to determine the extent to which geographic, social, and semantic distances are an explanatory independent variables for the observed frequency of contact between users. In order to examine this relationship, the operationalized geographic, social, and semantic variables are held as independent variables in a regression analysis where frequency of contact between users who tweeted is dependent. The only expectation is that lower geographic distance is a significant predictor, as has been asserted by Takhteyev et al. (2012). Beyond geography, social and semantic distance may provide further insight, and indeed, possibly outpace geography in terms of strength in the model. By employing these three variables, together, however, the objective is to show a more detailed and accurate picture of the importance of geography on Twitter.

Initial Data Omissions As mentioned, a significant portion of the sampled tweets were ultimately omitted from the study – of the 10,000 interactions, 1,299 were omitted due to either account deletion by the "responded" or "responding" user or deletion of either of the tweets involved in the interaction. Further, geolocated tweets are not all created equally – the provenance of geolocation data comes in two forms from Twitter's API, "Point" and "Polygon" data. Polygon data is a generalized rectangle drawn onto a map, indicating that the Tweet originated somewhere within that bounding box, while Point data is a precise, GPS-derived location of much higher accuracy, akin to a pin-point on a map. Upon review of the sample, it was found that many interactions had stated distances of 0 miles – the reason behind this being that two users employing "Polygon" data were located in the same bounding box, and as a result, the centroids of the bounding boxes for both tweets were identical. In effect, these interactions were not capable of accurately portraying meaningful distances between users, and were subsequently omitted in the analysis. Though this reduces the generalizability

of the research (in that GPS-derived data may carry many other assumptions, such as the use of more expensive devices capable of creating such data), it is a necessary omission due to its incapability to represent geographic distance. A total of 3,960 tweets were omitted in this process, further reducing the dataset to 4,741 observations.

Descriptive Statistics One of the major contributions of this work is its methodological distinction surrounding the unique role that GPS-derived data can play in delivering precise measurements of geography's salience. As such, a brief review of the basic properties of this data set may be an interesting point of discussion. The median log distance between interacting users was 2.69, translating to ≈ 14.2 miles. As the data is heavily skewed, the mean differs substantially, and was found to be 3.21, or ≈ 523 miles. Substantively, these findings are in accordance with the distance calculations found in previous work Liben-Nowell et al. (2005). The mean value for mutually-employed entities was 25.96, while the median was 6. Implicit triads were similarly skewed, albeit to a bit more severe of a degree – the mean number of implicit triads was 3.34, while the median was 0.

Figure 1: Scatter plot of total interactions by log distance between users at time of interaction. The high point, at 300, is a candidate for an outlier, but was ultimately left in the data as it was not a result of data error and was only twice as high as the next points, and uneven variance is expected.



Regression Modelling and Diagnostics As all data being observed is formally count data (disregarding geographic distance, which still may exhibit similar behaviors), the standard OLS regressions are largely an inappropriate approach for regression analysis, which is required to test significance of the independent variables. Count data is intrinsically heteroskedastic, and as such breaks a fundamental assumption of OLS modelling (Hilbe, 2011, 30). Even by transforming the data to likely candidates, the square-root and logarithm of the variables, Breusch-Pagan/Cook-Weisberg tests confirmed that the heteroskedasticity was significant ($\chi^2 = 1099.14, Prob > \chi^2 = 0.0$). A likely model candidate, then, is the Poisson regression, which is specifically built to model the typical attributes of count data – one fundamental assumption of this model, however, is that the data is not overdispersed – or that the variance of this data is not significantly higher than the mean. Two readily available tests, the Score test (Coef. 20.7, t-probability: 0.0) and Lagrange Multiplier test (LM value: 4.133e+67, P-value: 0.0), easily assert that the null hypothesis, that the data is equidispersed, can be safely rejected. Coupled with the test Poisson regression's Pearson dispersion statistic being relatively high (Pearson: 30.75472) and the number of operable observations being of sufficient scale (3,352), some distinct version of a negative binomial test is more appropriate (Hilbe, 2011, 179,183).

In assessing which specific derivation of negative binomial or Poisson regression is appropriate, the Akaike Information Criterion, coupled with comparative oberved-predicted difference graph of different models, the Bayesian Information Criterion, and Vuong tests, are particularly instructive. Again, the Poisson model, as well as the zero-inflated Poisson model, perform poorly – ultimately, zero-inflated negative binomial and negative binomial models prove to be most accurate, though the most successful is a zero-inflated negative binomial model, as Vuong's test reports a z-score of 2.43, corresponding to a p-value sufficient to reject the standard negative binomial model (0.0076). Ultimately, then, a zero-

Figure 2: Difference of observed against predicted values for variations of five candidate models. Although visually obscured, ZINB with both Probit and Logit characterizations of excess zeroes, as well as the NB2 model, produce similarly well-fitting results.

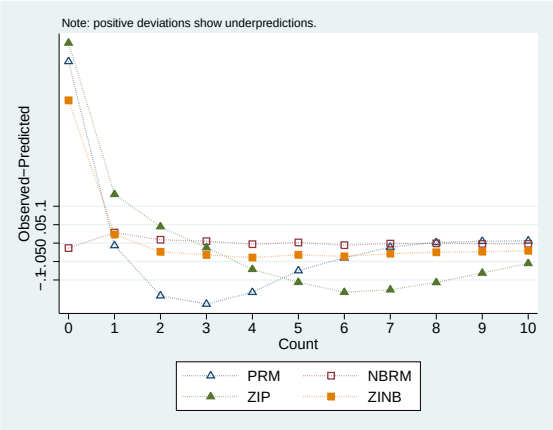


Figure 3: Output of Long & Freese's "countfit" operation, showing difference of observed values against predicted values for the four primary variants of Poisson regressions, including Poisson regression. While zero-inflated negative binomial regression is a poor predictor of values of zero, it is ultimately preferred over negative binomial regression through a review of AIC, BIC, and Vuong scores.

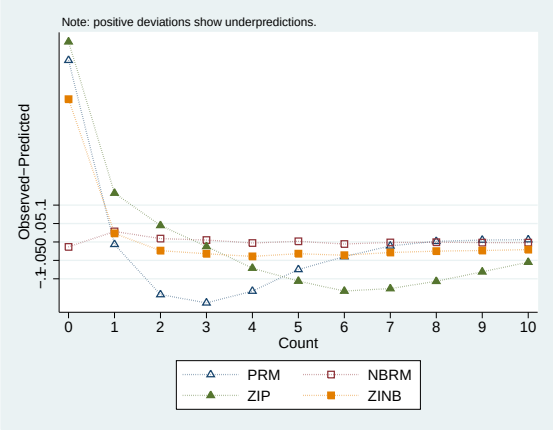


Figure 4: AIC and BIC scores for all tested models

	$NB2$	ZIP	$ZINB, Logit$	$ZINB, Probit$	$Poisson$
AIC	3.779056	8.101101	3.386415	3.779073	11.53797
BIC	3.785487	8.106255	3.400363	3.785504	11.53889

inflated negative binomial model with logit characterization of excessive zeroes is used. Substantively, the test asserts that some cases of zero interactions are "certain" zeros, while others are in fact over-represented due to some variable in the model – that expressed geographic, social, or semantic distance values significantly

correspond to zero-value observations (Long & Freese, 2006).

Outliers, Collinearity, Transformations As a zero-inflated negative binomial model is not capable of producing some of the more useful outlier detection methods, a more qualitative review of the data was employed in outlier detection for all variables. One outlier was found – one interaction falsely reported a value of 71,988 for the number of implicit triads (social distance) between the interacting users. This datapoint was investigated, and found to be a data error – the actual value was 1, and was corrected, though the provenance of this interaction was based on "Polygon" data, and was omitted from the regression. Collinearity diagnostics were run on the dataset, and through review of the SPLOM's as well as Stata's "coldiag2" function, no significant collinearity was found – lowess lines for the relationships for all independent variables tended to be relatively flat and unevenly distributed, and the highest condition index was 1.66, far below the minimum threshold of concern, 15. Finally, it should be noted that distance was transformed on a substantive basis as a log of distance as was employed by Liben-Nowell et al. (2005) – the log value of distance is a more reasonable approach to interpretations of distances between users, has been used in previous literature, and is thus employed here.

Model As is shown, all variables employed in the analysis are significant – the log of distance (z: -2.48, P > |z|: 0.013), number of implicit triads (z: 3.44, P > |—z|: 0.001), and number of mutually employed hashtags (z: 5.52, P > |z|: 0.000) all bear some explanatory power in the frequency of communication between users. Interestingly, however, geographic distance is not the best explanation for the variation seen in the frequency of contact between interacting users.

For every hashtag mutually employed by two users, the expected number of interactions between the users increases by $\exp(0.224) \approx 1.25$ times while holding all other variables in the model constant. As a result, the increase of mutually em-

Figure 5: Collinearity diagnostics results.

```
Zero-inflated negative binomial regression       Number of obs   =       3352
                                                 Nonzero obs     =       1527
                                                 Zero obs        =       1825

Inflation model = logit                          LR chi2(3)      =      75.92
Log likelihood  = -6342.484                      Prob > chi2     =     0.0000

------------------------------------------------------------------------------
total_inte~s |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
total_inte~s |
log_distance |  -.0374294   .0150803    -2.48   0.013    -.0669863   -.0078726
number~_ties |   .0087304   .0025353     3.44   0.001     .0037612    .0136996
number~ities |   .2235644   .0404979     5.52   0.000       .14419    .3029389
       _cons |     1.0583   .0754987    14.02   0.000     .9103249    1.206274
-------------+----------------------------------------------------------------
inflate      |
log_distance |   .4614362   .3196684     1.44   0.149    -.1651024    1.087975
number~_ties |   .0424935   .0187491     2.27   0.023      .005746    .0792411
number~ities |  -21.26826   12821.08    -0.00   0.999    -25150.13    25107.59
       _cons |  -7.977937   3.734751    -2.14   0.033    -15.29791   -.6579592
-------------+----------------------------------------------------------------
    /lnalpha |   1.522132   .0359972    42.28   0.000     1.451579    1.592685
-------------+----------------------------------------------------------------
       alpha |   4.581985   .1649384                      4.269852    4.916935
------------------------------------------------------------------------------a
```

ployed hashtags, or semantic distance, corresponds to an increase in the number of interactions users may have. While the data was shown to not be collinear, this result may be somewhat spurious. The problem arises in the mechanics of how the data was collected, and the nature in which the users employed the hashtags – if, for example, the two interacting users were directly responding to each other's tweets, it is likely that the same hashtags would be employed. One may imagine a hypothetical situation where User A notices User B's status talking about a particular subject, and it is categorized with a hashtag – User A's probability of employing that same hashtag would likely increase in the response they create, thus blurring a line between both the shown historical frequency of contact between users as well as the social distance between the users (substantively, a reasonable assumption would hold that the closer two users are, the more likely they are to engage in one another's content on the platform). The inflated portion of the model shows that the number of mutually employed hashtags is not a source of increased zero counts (STATA Corporation., n.d.).

Social distance is also a significant variable in this model, albeit to a lesser degree – for every mutual friend the two interacting users have (forming an implicit triad), the expected number of interactions between the users increases by

exp(0.0087304) ≈ 1.01 times while holding all other variables in the model constant. While this impact is minimal, the results are significant. Substantively, this means that compared to the log of distance or the number of mutually employed hashtags, the social proximity of the two users is not a particularly efficacious variable for predicting increases of interactions between users. One thing that is significant, though, is the impact that social proximity has on predicting certain zero cases. Again, the coefficient is relatively low, but significant – for every mutual friend between interacting users, the odds that their number of interactions were truly zero increases exp(0.042) ≈ 1.04 times. Substantively, this implies that more socially close users are slightly more likely to not have contacted each other previously (STATA Corporation., n.d.).

Geographic distance is significant, but only to a point. For every unit increase in the log of geographic distance, initially measured in miles, the number of interactions between users decreases by exp(-0.0374294) ≈ 0.96, or about a 4% drop. The inflated value was not significant. Substantively, this finding aligns itself with much of the previous literature: geographic distance matters. What is interesting, however, is the degree to which geographic distance matters when compared with other metrics, such as social and semantic distance. The interpretation gained from this model is that distance has relatively little impact on the frequency of interactions between users when considering other, more salient characteristics about the relationship between the two users. The series of assumptions and omissions the led to the regression that has been calculated make this result's interpretability very couched: this result applies to users who are active with other active users, who use GPS-enabled devices in their tweeting practices, and consistently post publicly. This population of users is likely different in many ways from other users on the service – GPS-enabled device use may imply many other things – purchasing the device and paying for the cost of a data plan for a GPS-enabled device (which would typically imply use of higher-end consumer

items such as iPhones, and is consistent with the applications users employed in these interactions) may translate to a fundamentally different type of user than one who simply uses the web interface on a computer immediately available to them (STATA Corporation., n.d.).

Conclusion

Geographic distance matters, but it is not the dominant factor. Mok et al. (2007)'s work focused on three components of analysis for understanding the salience of geographic distance for online interactions – the frequency of contact, the geographic distance between actors, and the underlying relationship between those actors. In this work, the underlying relationship was modeled as a combination of social and semantic distance, or who users bind to and what they talk about online. While geographic distance was found to be statistically significant, the strongest predictor of the frequency of interactions between users was the semantic component, or the things that people said to one another online, followed by the number of implicit triads, and finally, geography. While the results are significant, they are based on a series of necessary assumptions, omissions, and decisions that ultimately limit the degree to which these results can be generalized to Twitter as a whole, or even further to social interactions online. Most important, all the data employed was geo-located content which found its origins in GPS-enabled devices. While the degree to which GPS-enabled device-using users are significantly different in their habits and use of the platform is unknown, it should not be assumed that it is a minor difference. As such, the results are generalizable only to the degree that the discussion is related to those users, and even then, to users that consistently use the service, have not privatized their account, and talk to other users that express these same characteristics.

What is novel in this research, however, is the methodological distinction of us-

ing GPS-derived data. As Hale et al. (2012) point out in their work, self-reported location data is far from accurate, and the degree to which it varies is likely unevenly distributed. Further, self-reported location data is a slightly different form of information – it informs the researcher where a user identifies their principal location to be, not where they are in the moment that they are creating data and interacting with others, which may vary widely. By using this data, these concerns are completely addressed, and a new angle of insight is created in assessing the degree to which geographic distance still matters, specifically in the case of these users on Twitter, and more generally on the Internet. Alternatively, the purpose of undertaking such a study with this precise geographic data allows for a response to the prevailing literature on distance, most notably Takhteyev et al. (2012)'s work. By employing this internally valid GPS data, it is possible to further re-assert the importance of geography on Twitter, though to much less of a degree than Takhteyev et al. (2012) did in their assessment. Additionally, the underlying relationships employed in this work shows geography's place in the discussion. In turn, this helps us understand what Twitter is as a platform – the results from this work suggests that sampled users are predominantly semantically bound to one another; while they may create friendships with varied users, and may have some basic geographic decay in their friendships and communication, the most important aspect of their interactions is the nature of their content.

To be sure, further research is required. The single largest room for improvement in this niche of research done on the platform is the problem of generalizability. While no longitudinal numbers are known, it is likely that the proportion of tweets that are geotagged will increase in the future – while it will remain difficult to generalize these users to the population as a whole, a more substantial proportion of users who employ this service will reduce this problem. This study should be replicated at a point where the proportion has become sufficient, so as to leverage this methodological improvement to attain far-reaching, methodolog-

ically sound data on geographic distance's salience in online interaction. Beyond this concern, there are several others that have been noted – while collinearity was not found in this study, the connection between frequency of interaction and the number of mutually-employed hashtags may prove to be a significant issue in subsequent studies, and a more rigorous model should be employed, perhaps in the form of modifying either the measurement of semantic distance or the measurement of the frequency of interaction between users. Additionally, the sampling strategy for this research, while it was done necessarily to be generalizable to the population of tweets from which it came, may be augmented for a slightly different angle of research into this question: by sampling a random set of users, and then mapping out their interactions with other geolocating users, a very diverse data set may arise, where other salient variables about their various social contexts may be found. Ultimately, though, the work addresses a question central to Internet research – the degree to which technology is used to transcend physical space, and addresses it in a way that provides comparably more internally valid and novel insight to the question.

References

Anderson, C. (2008, June 23). *The End of Theory, Will the Data Deluge Make the Scientific Method Obsolete?* `http://www.wired.com/science/discoveries/magazine/16-07/pb_theory`. Wired.

Babbie, E. (2010). *The Practice of Social Research.* Belmont, CA: Wadswowrth Publishing Company.

Bar-Yossef, Z., Broder, A. Z., Kumar, R., & Tomkins, A. (2004, May 17-22,). Sic transit gloria telae: Towards an Understanding of the Web's Decay. In *Proceedings of the Thirteenth International Conference on the World Wide Web* (p. 328-337). New York City.

Bates, S. (1996, April 9-12,). The End of Geography. In *Symposium: Theories and Metaphors of Cyberspace.* Vienna, Austria.

boyd danah. (2007). Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life. *MacArthur Foundation Series on Digital Learning Youth, Identity, and Digital Media Volume*.

boyd danah, & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication Society*, 1-18. Available from `http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878`

boyd danah, Golder, S., & Lotan, G. (2010, January 6). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *HICSS-43. IEEE.* Kauai, HI.

Cairncross, F. (1997). *The Death of Distance: How the Communications Revolution Will Change Our Lives*. Harvard Business Press.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. (2010, May 23-26). Measuring User Inuence in Twitter: The Million Follower Fallacy. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (p. 10-17). Washington, DC.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2010). Who is Tweeting on Twitter: Human, Bot, or Cyborg? In *Proceedings of the 26th Annual Computer Security Applications Conference* (p. 21-30).

Daraganova, G., Pattison, P., Koskinen, J., Mitchell, B., Bill, A., Watts, M., et al. (2012). Networks and geography: Modelling community network structures as the outcome of both spatial and network processes. *Social Networks*, *34*(1), 6-17. Available from `http://www.sciencedirect.com/science/article/pii/S0378873310000614`

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American*

Society for Information Science, *41*(6), 391–407.

Fischer, C. (1982). *To Dwell Among Friends: Personal Networks in Town and City*. Chicago: University of Chicago Press.

Friedman, T. (2005). *The World is Flat: A Brief History of the Twenty-First Century*. Farrar, Straus and Giroux.

Fukuyama, F. (1989). The End of History? *The National Interest*, *Summer*, 3-18.

Golder, S., & Yardi, S. (2010, August 2022). Structural predictors of tie formation in Twitter: transitivity and mutuality. In *Proceedings of the second ieee international conference on social computing*. Minneapolis, MN.

Granovetter, M. S. (1973, May). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360-1380.

Hale, S., Gaffney, D., & Graham, M. (2012). Where in the world are you? Geolocation and language identification in Twitter. *Forthcoming*.

Hauff, C., & Houben, G.-J. (2011). Deriving Knowledge Profiles from Twitter. In C. Kloos, D. Gillet, R. Crespo Garca, F. Wild, & M. Wolpers (Eds.), *Towards ubiquitous learning* (Vol. 6964, p. 139-152). Springer Berlin / Heidelberg.

Hecht, B., Hong, L., Suh, B., & Chi, E. (2011). Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (p. 237-246). Vancouver, BC, Canada.

Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press.

Katz, E., & Lazarsfeld, P. (1955). *Personal Influence: The Part Played by People in the Flow of Mass Communications*. New York City, New York: The Free Press.

Kulshrestha, J., Kooti, F., Nikravesh, A., & Gummadi, K. P. (2012). Geographic Dissection of the Twitter Network. *Forthcoming*.

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* (p. 591-600). New York, New York.

Lewis, K. and Kaufman, J. and Gonzalez, M. and Wimmer, A., and Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, *30*(4), 330 - 342.

Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A. (2005). Geographic Routing in Social Networks. *Proceedings of National Academy of Sciences*, *102*(33), 11,623-11,628.

Long, J. S., & Freese, J. (2006). *Regression Models for Categorical Dependent Variables using Stata* (2nd ed.). Stata Press.

Lussier, J. T., & Chawla, N. V. (2011). Network Effects on Tweeting. In *Proceedings of the 14th international conference on Discovery science* (p. 209-220). Berlin.

Milgram, S. (1967). The Small World Problem. *Psychology Today*, *2*, 60-67.

Miller McPherson, L. S.-L., & Cook, J. M. (2001). Birds of A Feather: Homophily in Social Networks. *Annual Review of Sociology*, *78*(6), 415-444.

Mok, D., Wellman, B., & Basu, R. (2007). Did distance matter before the Internet? Interpersonal contact and support in the 1970s. *Social Networks*, *29*(1), 430-461.

Nissenbaum, H. (2004). Privacy as Contextual Integrity. *Washington Law Review*, *79*, 101-139.

Pennacchiotti, M., & Popescu, A.-M. (2011, August 21-24,). Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (p. 430-438). San Diego, CA.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010, October 30). Classifying Latent User Attributes in Twitter. In *SMUC '10.* Toronto, Ontario.

Scellato, S., Musolesi, M., Mascolo, C., & Latora, V. (2010). Distance Matters: Geo-social Metrics for Online Social Networks. In *Proceedings of the 3rd Workshop on Online Social Networks, WOSN 2010.* Boston, USA.

Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. (2011, July 17-21). Socio-Spatial Properties of Online Location-Based Social Networks. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (p. 329-336). Barcelona, Spain.

Scott, J. (1988). Social network analysis. *Sociology*, *22*(1), 109-127.

STATA Corporation. (n.d.). *Annotated Stata Output: Zero-Inflated Negative Binomial Regression.* UCLA Academic Technology Services. Available from `http://www.ats.ucla.edu/stat/stata/output/Stata_zinb.htm`

Stouffer, S. A. (1940). Intervening opportunities: a theory relating mobility and distance. *American Sociological Review*, *5*(6), 845-867.

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

Thrift, N. (1977). *An Introduction to Time Geography.* Geo Abstracts, University of East Anglia.

Tinati, R., Carr, L., & Tarrant, D. (2012, April 16-20,). Identifying Communicator Roles in Twitter. In *Proceedings of the 21st international conference companion on World Wide Web* (p. 1161-1168). Lyon, France.

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on twitter. In *Proceedings of the 20th international conference on World Wide Web* (pp. 705–714).