Spread Global, Start Local: Modeling Endemic Socio-Spatial Influence Networks

by

Jeffrey Solomon Block

A Thesis Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
MASTER OF SCIENCE
(GEOGRAPHIC INFORMATION SCIENCE AND TECHNOLOGY)

August 2015

DEDICATION

This thesis is dedicated to my mom and dad, who taught me to fight on.

ACKNOWLEDGMENTS

I would like to express gratitude to all of my professors at USC. The USC SSI faculty have served as sources of emulation throughout my time in their program. Special thanks to Dr. Swift for encouraging, and thus fostering, my interest in GIS development, and Dr. Kemp, who has elevated my appreciation of expert tutelage to a level I did not anticipate. To my colleagues, many thanks for stoking the flames of my competitive spirit, and understanding when I had to bow out from time to time and finish this monster. Most of all, I would like to thank my family, specifically my lovely wife—this thesis would not have been possible without her tireless support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AOI	Area of Interest
CSV	Comma Separated Values
ESSLVM	Endemic Socio-Spatial Latent Variable Modeling
GUI	Graphic User Interface
ISIL	Islamic State of Iraq and the Levant
IVO	In the Vicinity of
LA	Los Angeles
ROC	Receiver Operating Characteristic
SSI	(USC) Spatial Sciences Institute
USC	University of Southern California

ABSTRACT

The importance of social media-borne influence has been demonstrated in dramatic fashion on a global stage, with examples ranging from the regime toppling Arab Spring between 2010 and 2012, to the startling ascendency of ISIL in 2014. The value of this influence however, is highly versatile in application, and not limited to geopolitics. Commercial marketing campaigns hinge on the propagation of their message through social networks, and social media influence practitioners have engineered methods of ensuring optimal results. This practice however, is often conducted solely in a virtual environment, where false positives can abound due to disconnection from geospatial ground truths. I have outlined a system to reduce network uncertainty and identify key influencers in a manner that improves upon existing analytic processes by geospatially decomposing nebulous social media networks into locally relevant networks, wherein tangible results are more likely. This study introduces a novel approach, demonstrating that position in a social network has bearing on an individual's relationship with others in physical space, and as a result, individuals or organizations postured to influence a network via direct conduits such as local leadership figures and on-site organizers, possess a qualitative advantage. Additionally, because there exists a reciprocal relationship between an individual's position in a social network and their position among others in physical space, geospatial assessment techniques can be used to infer social connections. Dubbed endemic socio-spatial latent variable modeling (ESSLVM), this method has been automated as a Python tool that can be integrated into ArcGIS. Concepts are demonstrated using a Twitter dataset from the late-November 2014 protests in Ferguson, Missouri.

CHAPTER 1: INTRODUCTION

> *You see how an idea spreads and becomes a worldview, and how the bearer, the individual, reaches out to form a community, and how an organization, then a movement, grows from the individual. The idea is no longer buried in the heart and mind of an individual. Now there are four, five, ten, twenty, thirty, fifty, eighty, a hundred, and ever more. That is the secret of ideas; they are like a wildfire that cannot be restrained. They are like a gas that seeps through everything. Where an idea finds entry, it enters, and soon that person is influencing others. (*Goebbels 1934, 3)

Influence is defined as the ability to engender some manner of change without resorting to a direct action expressly intended to achieve the same effect—e.g. physically coercion as a means of dissuading someone from continuing a certain behavior does not fit the classical definition of influence, though in the process you could influence others (merriam-webster.com). Throughout this study, influence is examined in a social network context, where it is defined more precisely as the ability to affect diffusion, or the spread and adoption, of ideas and practices (Kadushin 2012).

Social media as a platform for the projection of influence benefits from an ability to assume the form of its user base and their prevailing narratives—an implement of the masses. This benefit however, is also often a bane. Where some see an opportunity to kick-start a charity, implement social change, promulgate a fashion trend, or market a product; others bully, conspire, and use ghoulish propaganda to raise armies of zealots and fanatics.

1.1 Influencing Social Movements and Philanthropy

In terms of social change, the medium has given the disenfranchised a voice otherwise denied to them by oppressive regimes. Social media has played a notable role in most recent socio-political movements (Cole 2014). Using Twitter and other social media during 2009 Iranian elections, the dissident Green Movement was able to circumvent state controlled media and broadcast its message, especially via mobile platforms that enabled dynamic reporting on breaking events (Khonsari et al. 2010).

Assessing the Green Movement's key influencers from a social network standpoint, the University of Tehran asserted that despite the network's apparent resilience, "there exists a strongly connected active core with large centrality metrics. This core is responsible for induction of information into the network. Since there are few actors in this core, it is possible to manipulate the knowledge state of the social network by controlling them" (Khonsari et al. 2010, 415). Individuals galvanizing a support-base through the spread of small-format media is not unfamiliar to Iran's ruling elite. As a prelude to the 1979 Iranian Revolution, the then exiled Ayatollah Khomeini distributed audio cassette tapes of his sermons throughout Iran and worldwide as a means of sowing influence (Sreberny et al. 1994).

Other key examples of social media's role in social movements include but are not limited to 2010-2012 events throughout the Arab Spring (Taylor 2012), and the "Euromaidan" revolution that resulted in the February 2014 overthrow of the incumbent Ukrainian regime (Bohdanova 2014). As testament to its effect, authoritarian countries coping with restiveness have taken to shutting down social media services; a countermeasure aimed at denying influencers a means of propagating their message (Cole 2014).

On the theme of positive change, charities have also availed themselves of the benefits of social media. The charity Feed America increased web traffic by 250% using a social media campaign, and Esri's actions in support of recovery efforts after the 2011 Japan earthquake were significantly improved by enabling social media within a web application (Twitter 2015). If you've ever changed your profile picture to stand up for someone's civil rights, shared content to raise awareness, or accepted a dare in the name of charity, you've been in the path of social influence. The Ice Bucket Challenge for example, was a charitable movement intended to support the ALS Association through interconnected acts of participation. It was a viral social media phenomenon that reached millions of people and raised in excess of 100 million dollars (Skarda 2014). These movements however, don't permeate through social networks at a constant pace, they are pushed through by influencers (Cha et al. 2010). A better understanding of this process could allow charities to achieve more with less.

1.2 Social Influence in Business

From a commercial standpoint, social media marketing spending is expected to rise 23% between 2014 and 2019—from 7.52 to 17.34 billion dollars (Statista 2015). It's a steep increase that is in part attributed to the growing trend of influence marketing, a concept that targets individual advocates who are likely to appeal to market segments, instead of directly marketing to a broader audience (Wong 2014). According to Forbes, influencer marketing has proven to generate higher sales and retention rates than traditional paid advertising (Wong 2014). Underscoring the efficacy of influencer engagement as a marketing modality, a major fast food chain simply started casual discussions with Twitter accounts followed by more than 10,000 other accounts, and in doing so, saw their online popularity vaulted to three times that of a chief rival (Elliot 2014).

Looking to capitalize on this approach, the market is crowded with social media influence brokers. Each however, applies their own formula to the influence quandary, and with varying degrees of success. In a case study on invigorating influencers, social media manager Hootsuite asserted "reach is not the same as influence" (Hootsuite 2014) to denote that value cannot necessarily be achieved through volume. As this thesis emphasizes, comprehensive data analysis across different social network variables has proven critical to finding true value in influence assessment (Bakshy et al. 2011). Reviews of other social media analysis services make it clear that there is a premium on message resonance (Notess 2013; Internet Wire 2013).

1.3 Social Media and International Terrorism

Different forms of social media also allow for malign actors to propagate harmful information, as evidenced by social media's new grim connection to international terrorist groups (Maher 2014). A study on ISIL by the Brookings Institute underscores the extent to which the group has parlayed social media to project influence disproportionate to its numerical strength; posting official statements interspersed with summary executions, including lurid acts of decapitation and immolation (Berger et al. 2015). According to the Brookings Institute study, despite routinely having their accounts shut down by Twitter, the group is still able to get ahead of restrictions by working through a multitude of "swarm" accounts to promulgate their message. Where countering this scourge is concerned, on the whole, account suspensions have proven ineffective at attenuating the group's global influence.

Identification of influencers in the path of such nefarious social media transmissions could be key to preventing wide and unchecked dissemination of malign or misleading content to audiences susceptible to Islamic (or religious) radicalization. Shedding light on the support base for this message, activist group Anonymous has revealed thousands of supporting ISIL Twitter accounts (Cuthbertson 2015).

Remarking on interest in social media, the commander of United States Operations Command (USSOCOM), Army Gen. Joseph Votel expressed interest in approaching social media as another facet in an unconventional warfare strategy. In particular, Gen. Votel emphasized the need to use social media to engage key influencers in order to effectively respond to unfolding crises (Gertz 2015). Often centered on influencers, defined social clusters have spatiotemporal properties that are assessed to play a key role in the dynamics of larger networks (Lucente et al. 2013). Consequently, there stands a requirement for an approach to social network analysis that will allow for a better understanding of spatiotemporal variables that can mitigate the spread of malign content by identifying key influencers among these social clusters and tailoring information campaigns accordingly (DARPA, 2014).

1.4 Outline of the Proposed Model

Social media borrows some of its lexicon from epidemiology. If something becomes wildly popular, it is considered "viral." The diffusion and acceptance of ideas in a social network is also referred to as "contagion." In this sense, influencer marketing and engagement is akin to the bio-warfare modus operandi of targeting transportation hubs to make the greatest use of a limited amount of pathogen. In such a context, this study focuses on local networks and their endemic phenomena, applying a process henceforth referred to as endemic socio-spatial latent variable modeling (ESSLV).

A latent variable model uses known variables to identify other variables that exist, but are otherwise unobserved. You've likely already been exposed to comparable methods in some form. Anticipatory advertising assesses what you're likely to buy based on what it knows you've bought, and what you're browsing for. Courtesy of artificial intelligence, these anticipatory methods are growing increasingly incisive, and they don't intend to stop at advertising (Thomas 2013). From a social network standpoint, there is Facebook's attempt to link you with other accounts who share your connections, a service referred to as, "People You May Know." The latent variable is the undocumented relationship between a member account and other member's accounts Facebook suggests (Klee 2014). Facebook also makes this data available through their Graph API (Facebook 2015). Ostensibly advancing their latent variable craft, in 2012 Facebook acquired mobile application Glancee, an app that juxtaposes its user-base's physical locations and interests to achieve social discovery, or the identification of ambient people and places that might appeal to a given user. Following its acquisition by Facebook, Glancee was taken offline, but it stands to reason that their technology is being assimilated (MacManus 2012).

Such methods serve to reduce network uncertainty, or the existence of unknown social network variables (Fuhrt 2010). Since influence in the context of social network analysis is characterized as one's effect on the propagation of information, understanding the dynamics of relationships along which information is passed is integral to the assessment of influence variables. As a pretext of this thesis, and as demonstrated with the Glancee example, it is assessed that spatially enabled latent variable models can be more insightful than those bereft spatial content, especially at the local level. Lending credence to the relationship between social and spatial (socio-spatial) variables, there exists numerous other models of this type covered in Chapter 2.

The key hypothesis of this study is that *position in a social network has bearing on an individual's relationship with others in physical space, and as a result, individuals or organizations postured to influence a network via direct conduits such as local leadership* *figures and on-site organizers, possess a qualitative advantage.* An ancillary assertion is that because there exists a reciprocal relationship between an individual's position in a social network and their position among others in physical space, geospatial assessment techniques can be used to infer social connections. ESSLVM, is a heuristic approach that offers novel spatial methods for optimizing local influencer identification. A case study, here referred to as a vignette, using social media events collected during the late 2014 Ferguson, Missouri protests, demonstrates the significance of augmenting known social relationships with GIS derived spatial connections, in order to diminish network uncertainty and better identify influencers.

This study explores the aforementioned analytic imperatives through the following research objectives applied to the vignette's area of interest:

- *Identification of communities, or social clusters, using geospatial correlation methods.* Through automated geoprocessing functions, social media events represented by a point vector dataset are binned by spatial correlation to a defined area or areas of interest within a larger extent. This process enables all social media entities whose users are active within the areas of interest to be interrogated and subsequently identified when their users appear elsewhere across the complete extent.
- *Querying the Twitter Friends and Followers API for unilateral and reciprocal social media contacts (friends and followers) of accounts identified from within social clusters.* This data is acquired via a Python script and converted into an association table for ingestion into a social network analysis application.
- *Isolation of the local social network from the complete social network, and the development of socio-spatial metadata.* This includes reduction of the network per

spatial criteria, in order to isolate locally significant social media accounts. Scores derived from the social network analysis of those accounts is appended to the point vector social media dataset of events for subsequent socio-spatial analysis.

- *Identification of socio-spatial relationships*. Using a Python script developed for this analysis, social media account events represented by the point vector dataset are assessed for social relationships based on the time difference between events, their spatial proximity, their social network proximity, and the number of times all user-specified criteria thresholds are met.
- *Analysis of the correlation between account social network metrics and the geospatial distribution of the related social media events.* This entails development and use of a Python script that employs Spearman's Rank Correlation Coefficient to assess the correlation between social and spatial variables on an individual basis.

Beyond this study's vignette, a menu of practical applications options for ESSLVM could include charitable outreach associated with a humanitarian event, an assessment of market influencers associated with a particular facility, or the development of obscure illicit networks in a specified location.

This study uses commercially available social media influence metrics, social network analysis software, organically developed scripts, and insight derived from a comprehensive review of existing latent variable models, to develop the ESSLVM system and demonstrate its value using real-world social network event data. The efficacy of this model was tested by conducting subjective research on top social media accounts associated with the Ferguson, Missouri real-world vignette. Included research entailed a characterization of key accounts' contributions to the vignette as posted in other public social media outlets, and a comparison of this information to the model's output. Ultimately it is concluded that the application of spatial analysis to social network analysis can be used to identify local influencers more effectively than social network analysis alone, and that specific methods introduced by ESSLVM are conceptually unique.

CHAPTER 2: REVIEW OF SPATIALLY AUGMENTED SOCIAL NETWORK ANALYTIC METHODS

As an interdisciplinary union, it is asserted that social network analysis augmented by GIS can offer demonstrative advantages over social network analysis conducted in a geospatial vacuum. Benefits however, can also be synergistic, with social network data contributing to geospatial insights—as demonstrated by latent spatial variable models covered in this chapter.

The pairing of social and geospatial analyses however, requires understanding of the relationship between their unique variables, as well as the data needed to conduct analysis in both disciplines together. A key concept addressed here is that of propinquity effect, which posits that entities near one another are more likely to demonstrate similar behavioral characteristics and form relationships (Bonito et al. 2002).

To what extent is propinquity effect applicable to contemporary social media datasets, and at what scale? A review of available literature demonstrates a reciprocal relationship between social and spatial variables across a range of vignettes. This research serves to explore the theoretical basis of endemic socio-spatial latent variable modeling (ESSLVM).

2.1 Twitter as a Data Source

The central component of a Twitter micro-blog, or tweet, is the 140 maximum characters that each tweet contains, allowing an account holder to express themselves via these staccato transmissions, which are posted to their Twitter account page, or into the Twitter feeds of those following them (Twitter 2015). The fine mechanics of Twitter however, include a great deal more. While all Twitter functionality is not addressed here, key Twitter modes of interaction and metadata fields used in this study's socio-spatial applications are enumerated.

An example of a visible tweet, that which is posted, includes several components that are indicated by numbers in Figure 1 below. The components are the user's profile picture or avatar (1), user name (2), screen name (3), the time and date the tweet was posted (4), and the body of the tweet (5). Additional information can include a user mention, or the citing of another screen name in your tweet (6), and a key word or phrase marked as a hashtag (7). Key modes of interaction with this tweet are indicated by replies to it (8), the reposting of the tweet or a "retweet" (9), as well as the designation of the tweet as a favorite (10). Other options are available via "More" (11), and include the ability to share a tweet via direct message, copy its link, and embed the tweet, among other choices. There is also indication of the tweet's spatial enrichment (12) as specified by the user or through geo-tagging. A geo-tagged tweet is as spatially accurate as the smartphone used to broadcast it, which in the case of contemporary smartphones can be within three meters, 90% of the time (Shaner 2013). The tweet depicted in the figure has also been augmented through an aftermarket application, with an influence score (13) courtesy of the Klout social media influence metric—which is subsequently addressed in this chapter.

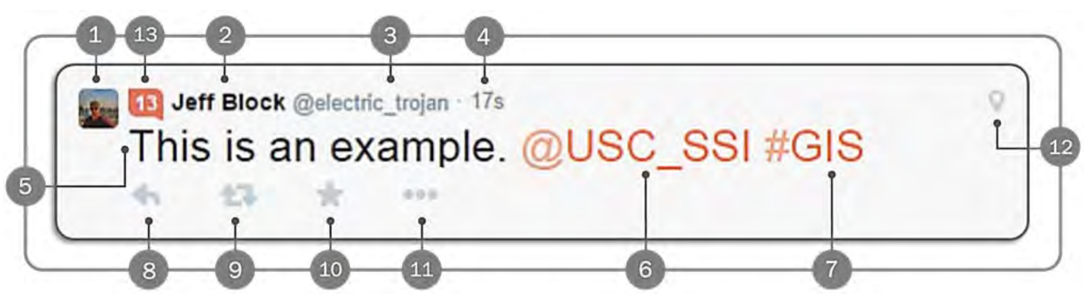

Figure 1 the Anatomy of a Tweet

For every tweet however, there is an abundance of additional information stored as metadata. This can include everything from the URLs associated with profile photos, to the tweet's language. Some fields are nested, and as such their quantity is dependent on volume of content, e.g. for each additional user mention or hashtag, there is an additional field. All metadata was not used in this study, however those fields maintained through all phases of the study are listed in Table 1.

Table 1 Twitter metadata used in this study

Field	Description	Data Type
Created_At	The date that the tweet was created	String
Text	The actual body of tweet, or its unstructured text	Text
In_Reply_to_Screen_Name	If a tweet is in response to a tweet from another account, the addressee's screen name is included here.	Text
User_ID	A unique numerical identifier for the Twitter account	Integer
User__Protected	Content not available publically. All accounts used in this study were not protected.	Binary
User__Followers_Count	The number of unilateral relationships are directed at the specified Twitter account	Integer
User__Friends_Count	The number of reciprocal relationships an account has with other accounts	Integer
User__Listed_Count	The number of Twitter lists that include the account. Twitter lists include a designated group of Twitter users, as codified by other accounts	Integer
User__Statuses_Count	The total number of tweets from an account, including retweets	Integer
User__Geo_Enabled	Whether or not the user has manually enabled location data, or the geotagging of their tweets	Binary
Geo__Type	Using only geotagged tweets, all Geo_Type was "Point"	Text
Geo__Coordinates001	Lattitude of geotagged tweet	Float
Geo__Coordinates002	Longitude of geotagged tweet	Float
Place__Full_Name	A place designated by the user, or assigned per the location of their geotagged tweet. The Place category also includes a littany of other spatially descriptive fields	Text
Entities_User_Mentions	Other screen names included in the body of the tweet	Text
Entities_Hashtags__Text	All hashtags, or key words and phrases preceded by the "#" character and devoid any spaces	Text
Entities_URLs	URLs included in the tweet	String

2.2 Key Socio-Spatial Concepts

This section establishes a social network lexicon and introduces social network components most applicable to geospatial analysis. With identification of influencers, a key research objective, the nature of influence is explored from a social network perspective. This includes how an individual's position in a social network has bearing on the influence they wield, and how understanding their relationship with others in the network can allow for the quantification of certain influence variables. Most importantly, social network analysis variables are explored as forces ultimately beholden to geospatial dynamics, even in an age of virtual communication.

2.2.1. Social Network Analysis

A social network, or social graph, is an abstract depiction of social relationships, with each relationship represented in its most basic form as a dyad, or a pairing of two nodes or vertices (Kadushin 2012). In framing this concept as an influence network, think of those who influence you, members of your family, colleagues, teachers, celebrities, and politicians—thought leaders of any kind. In this social graph, these people are nodes, and your connections to them are edges, links radiating out from you—the principal node (Kadushin 2012). Ask your influencers about their connections, and your known network expands. No longer just the hub and spoke configuration of you and your immediate influencers, from this multi-tier network, vicarious connections emerge, and transmission of influence through intermediary nodes becomes evident, as illustrated in Figure 2. To what extent are ideas and influence transmitted beyond known contacts?

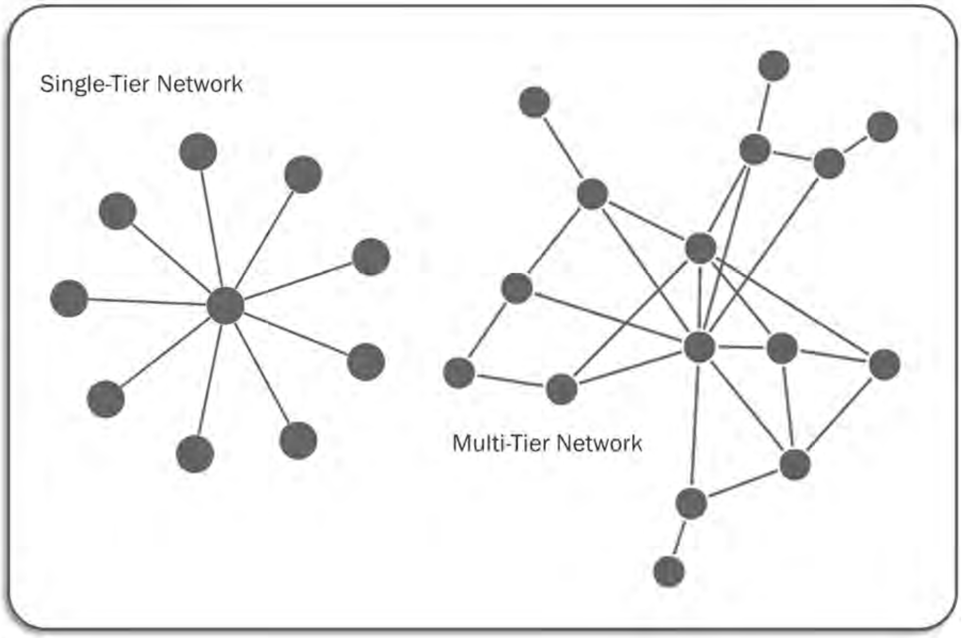

Figure 2 Single and Multi-Tier Social Network Graphs

Where extended social ties have been assessed as a vector for influence, studies on behavioral homogeneity demonstrate that connections out to a third tier consistently propagate meaningful influence through a network (Christakis et al. 2011). The propagation of content through a network is characterized as information cascade, or the aforementioned diffusion. To elucidate this process, consider earlier examples of "viral" social media content fomenting a response as it spreads to millions. The documentation of these cascades as concatenated social media events, such as those employed in viral campaigns, has revealed how significant individual nodes are to the projection of influence beyond immediate personal associations (Hodas et al. 2014), with select nodes wielding disproportionate influence (Cha et al. 2010). Consequently, variables such as a node's position in an association network, or its centrality, are key to governing the passage of information and thus a fundamental element of social network analysis (McCulloh 2007). In addition to the value of centrality towards propagating the spread of influence through a network, identification of key nodes based on certain centrality measures can inform efforts to either intentionally fragment a network, or prevent this type of network disaggregation from occurring. Basic forms of centrality—degree, closeness, and betweenness—are subsequently addressed and depicted in Figure 3.

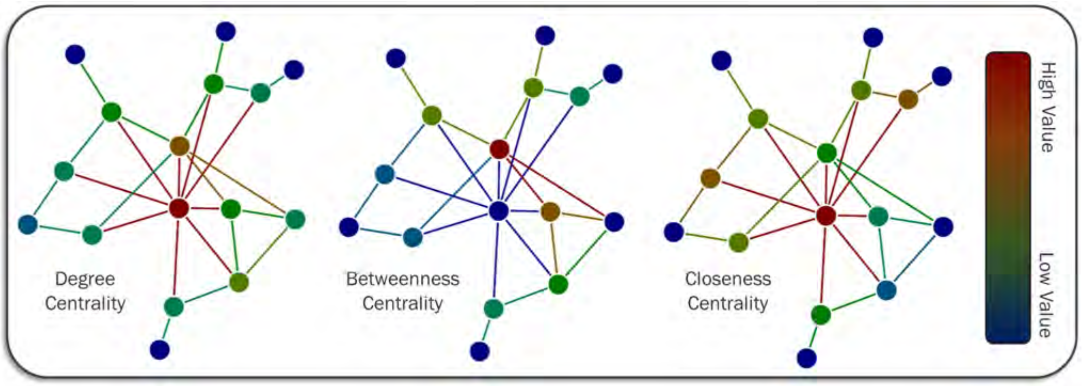

Figure 3 Examples of Degree Centrality (Left), Betweenness Centrality (Middle), and Closeness Centrality (Right)

The most basic form of centrality is degree centrality, the number of total edges, or relationships that a node has (Wasserman et al. 1994). A node with high degree centrality has the potential to be most aware of events in their portion of the network based on its high level connectivity. Conversely, high degree centrality means that a node has a greater opportunity to directly project influence to those immediately around them.

Closeness centrality is a measurement of the passage of information from a node to all others in the network (Freeman 1979). Movement from one node to another is considered a step, and closeness centrality is determined by identifying the shortest path from one node to every other node in the network, as a measurement of steps, and using an inverse of that average to calculate a centrality value. While a node with high closeness centrality might not have the connections of a node with high degree centrality, it would be better postured to be cognizant of goings on throughout the entire network, and could affect the spread of information most efficiently.

Betweenness centrality is derived from the shortest path measurement used to determine closeness centrality, except betweenness centrality is a measurement of the number of total shortest paths in a network that intersect a particular node (Freeman 1979). Consider this to be an indication of which nodes sit along the most important positions on major network thoroughfares. As nodes along the shortest paths between other nodes, entities with high betweenness centrality can bind otherwise disparate network clusters, and serve to control the spread of influence throughout the network.

2.2.2. Semantics of Node Characterization in this Study

In this study, where actors transcend multiple media there is a requirement for clearly articulating how these flesh and blood actors as well as their virtual representations are characterized. While such characterizations are not externally definitive, in this this study they are as follows:

- *Entity* refers to either an individual or their social media account, if that social media account is serving as a surrogate for a presumed sentient being, mobile or otherwise. Entity is preferred over "person" as even when geotagged, this data only constitutes a virtual representation. *An entity can have many locations in space, but only one in the social network*.
- *Account* can be used interchangeably with entity, but only as it pertains specifically to social media.
- *Node* is graph lexicon, and as such refers only to an entity's position in a social network, not physical space.

2.2.3. Directionality and Information Diffusion within Social Media Networks

When modeling influence, the passage of information through a network is critical to isolating influence hubs (Pew Internet & American Life Project 2014). The transfer of information however, is subject to the directionality of the ties over which it passes (Hoi 2011). This is essential to distinguishing between nodes such as broadcast outlets—characterized by unilateral connections—from those involved with their networks (Pew Internet & American Life Project 2014). In graph parlance, where an edge is a connection between two nodes or vertices, an arc is a directed connection which can be unilateral or reciprocal. This is especially important when using both Twitter friends (reciprocal) and followers (unilateral). From a social network standpoint, edges leading to a node constitute "in degree" centrality, and those leading from it would be "out degree" centrality (Wasserman et al. 1994). A social network score based on both forms of degree is total degree centrality. Accordingly, a Twitter account with many followers and few friends would have high in degree centrality, and low out degree centrality. The directionality of friends and follower relationships are leveraged in this study, as the use of both arcs and edges in social media analysis has proven advantages over the use of undirected social graphs (Sadinle 2013).

Within social media networks, and in this case Twitter, methods of assessing information diffusion include multiple modes of documenting a social media event's resonance. Among these modes is tracking the propagation of retweets through a network. Since a retweet is the reposting of one user's content by another, a cascade of such reposts can spread along a chain of connected accounts, with each account actively passing the message on to the next. Taken into consideration with other centrality measures, retweets have proven to be a key indicator of network influence, per information diffusion models (Anger et al. 2011).

A different approach uses the specific mention of one account screen name in another account's tweet. As addressed in the Twitter metadata section, this is referred to as a user mention, and it can be used as an indication of the significance of the mentioned account to others in its network (Cha et al. 2010). Where this type of interaction serves as a metric for social media influence, the total number of tweets generated by a user can be used to normalize the number of times their content is retweeted or they are cited via user mention.

2.2.4. Why Social Network Popularity is Not Influence

If influence in a social network context is the ability to affect diffusion, or the spread and adoption of ideas and practices (Kadushin 2012), and a node exhibits high levels of degree centrality, is that node more influential by virtue of the sheer number of other nodes with whom it shares a relationship?

In attempting to quantify influence on Twitter, Bakshy found that an entity's gross number of followers was unsurprisingly directly correlated to influence metrics (Bakshy 2011). Intuitively, the more followers an entity has, the more influential it is likely to be. In Bakshy's study, influence was determined by using the documented occurrence of information diffusion events that occurred as retweets. However, from a standpoint of value, in a scenario where an entity's total number of followers are tantamount to the cost it would take to employ them as an advocate—per the assumed proportionality of degree centrality to influence—engaging the most followed influencers was not as effective as engaging multiple average influencers.

Where both followers and directional ties are both used to assess influence, in a study by Saito (2013) that compared two Twitter accounts, both with an equal amount of followers, but one with significantly more reciprocal relationships, the account with a superior number of reciprocal connections achieved a higher tweet to retweet ratio, and thus was perceived to be more influential.

A different study addressed the nuance of a purported correlation between total follower count and influence (Cha et al. 2010). Using retweet and user mention ratio as a surrogate for influence, the Cheng study determined that while correlation coefficients between follower count and information diffusion values were passable as applied to all accounts, when restricted to the top 10% of those exhibiting strong influence characteristics, follower count performed poorly as an explanatory variable. This study also concluded that those accounts that did achieve a disproportionately high level of diffusion had cultivated their standing as an influencer through sustained engagement. These findings compete with the Bakshy study in the sense that they are at odds with assertions about the value of follower counts. However, flouting follower counts as an influence fallacy does, to an extent, appear to lend credence to Bakshy's findings on the value of engagement at lower levels of the network.

As a result of research such as that cited in the Bakshy study (Bakshy 2011), commercial literature on social media campaigning encourages reaching out to certain influencer types at various levels. The public relations firm Augure separated influencers into three distinct classes (Augure, 2014). At the highest end of the popularity spectrum are celebrities, assessed as ideal for sponsorship opportunities, followed by opinion leaders or segment experts, and finally consumer influencers, who are best postured to support marketing campaigns. The Augure study laid out three key components of influencing vis-à-vis virtual networks, these included an account's exposure or reach, level of participation within their community, and the account's "echo," or the likelihood that content they generate will be relayed through the network—which is consonant with retweet and user mention models.

On the whole, it has been asserted that identifying influencers using social network analysis requires additional information on the interplay between nodes. Where this information is available, it is assessed that choosing on numerical superiority, from a degree centrality standpoint, is not always an optimal influencer engagement strategy. Additionally, one must design an engagement strategy around an intended outcome, and seek out a corresponding influencer type. Engagement at lower levels, both socially and likely geospatially, may offer advantages that incorporate these methods.

2.2.5. Social Networks to Assess Communities and their Influencers

In isolating influencers at lower tiers, analysis of local networks necessitates parsing of clusters into social communities. This community detection concerns identification of structures within a social network that can be grouped distinctly from the rest of the network (Guo 2009). This is especially relevant to the study of the Twitter users, which consists of numerous distinct clusters suspended in an otherwise loose overall network (Macskassy 2012).

Community detection is also relevant to anomaly detection through spatiotemporal clustering. This method takes into account community attributes to identify trending events by comparing current social media levels to endemic community baselines (Pozdnoukhov et al. 2011). Thus, knowledge of a community's physical and social dimensions affords an analyst the ability to assess factors most capable of impacting the network. To this end, by using network clustering algorithms and ground truth assessments to identify communities (Yang et al. 2015), significant improvement was noted over traditional virtual assessment methods where there was no access to the study area afforded.

To assess the effects of social communities in a virtual medium, such as Twitter, on physically distributed communities, geospatial tweet density has been linked to tangible socioeconomic factors (Lia 2013). Additionally, Mennis (2011) demonstrated that the location of social network nodes has an impact on associated localized sociocultural behaviors. With this in mind, the effective partitioning of geographically local communities within a social network can allow other network influence metrics to be more effectively brought to bear, and social engagement strategies to be more judiciously implemented.

2.2.6. Spatial Mobility as a Social Network Variable

When identifying communities per a spatial extent, the mobility patterns of those that comprise the community are key. Mobility studies reviewed for this study (such as Becker et al. 2013) focus on the use of mobile devices—the principal means of transmitting a geotagged tweet. These studies, however, have been conducted using catalogued or historically inferred social network spatial patterns, e.g. assessed human mobility behavior in the 1944 Budapest Ghetto (Giordano et al. 2011). Human mobility patterns are not an intended contribution of this study, however research from the Becker study was applied to the Ferguson, Missouri vignette in order to approximate a mobility extent that can adequately circumscribe relevant agents.

Within the study reported here, mobility patterns as they apply to direct and reciprocal social dynamics were measured using Euclidean distance, with distances out to 50 meters used as a parameter for prospective relationships. Previous research indicates this is a reasonable consideration for the occurrence of transitory social contact (see for example, Brieger et al. 2003). Within this distance, anthropogenic features that would otherwise impede or obstruct social behaviors, such as a road network and its trafficability, are assumed to not be an issue.

2.2.6. The Spatial Implications of Social Network Analysis

With social network centrality concepts addressed, now it is possible to consider what the spatial component's contribution to forms of social network and latent variable analyses is. It was introduced as a means of improving social network analytic methods, but how, and to what extent has this been substantiated? Furthermore, virtual communication has changed our concept of space in the context of relationships. Are concepts that address social network relationships on a physical plane applicable to virtual realms?

Using Twitter data to define edges in a social network, it was observed that social media connections are more likely to form between nodes as a result of some form of shared geospatial correlation (Takhteyev 2013). Statistically, it has been shown that, even in vast virtual networks, nodes are more likely to impart more significant influence as a result of physical proximity to other nodes (Sagl et al. 2014; Bonito et al. 2002).

As a function of propinquity effect, several studies have indicated that the concept of social network uncertainty is likely to be significantly reduced where the distribution of nodes in physical space is known or even approximated—endowing distance functions with the ability to predict social network structures (Butts 2003; Daraganova et al. 2012).

2.3 Social Media and Social Network Applications

Having described the commercial significance of social media influence analysis, it should come as no surprise that there is a veritable cavalcade of social media analysis application offerings on the market. These can range from free online applications with premium upgrade options, to comprehensive social media analysis software as a service. In all forms, social media data is analyzed or otherwise manipulated in order to offer some manner of analytic insight.

This section analyzes three classes of social network tools: social media influence metrics as an online service, social media analysis as a server-side application and subscription service, as well as social network analysis client-side software that is not inherently social media-based.

2.3.1. Social Media Influence Metrics as an Online Service

Numerous companies offer social media analytics as a commercial service. These services typically consist of a web interface that allows users to input social media data in the form of accounts or topics that are rated in some fashion—most often to determine influence. These services are plentiful, and the means by which they extrapolate influence metrics is typically proprietary, and advertised as multi-faceted but explained in basic terms (Notess 2013).

As an example, in the case of industry leader Klout, early attempts to master the influence formula elicited a critical response when scores favoring raw connections, or degree centrality, favored a major pop star over the President of the United States (Internet Wire, 2013). It is worth noting however, that Klout has recalibrated its algorithm to account for more network resonance cues, and remains at the forefront of social media influence metrics (McHugh 2012). Such issues however, speak to the nuance associated with assessing influence, especially across heterogeneous mediums—namely virtual and actual. The following sampling of services were reviewed for influence assessment functionality:

- Klout makes use of a proprietary algorithm that draws upon information aggregated across different social media sources. Their method takes into account 400 different signals to provide a score on a scale of 1 to 100 (McHugh 2012). These signals include various means of measuring an account's network size and level of interaction therein. It is worth noting, that other services listed below will often include a Klout score in addition to their own influence metric.
- Tellagence, which is characterized as a social prediction company, conducts influence and trend analysis. Tellagence has promoted itself, at Klout's expense, as employing more holistic methods focused on the flow of information through a network instead of metadata values (Internet Wire 2013).
- SocialRank (socialrank.com) allows the user to sort accounts based on location, keyword and location, among other filters. Ostensibly, this service assigns scores in part associated with an account's level of interaction with followers and those they are following vis-à-vis their tweet volume.
- MoKumax offers a service known as Twitter Grader. This service provides an influence score ostensibly derived from tweet volume, retweets, replies to others and the number of times a user has been retweeted—potentially among other contributing factors. MoKumax also generates exceptional visualizations based on tweet characteristics, such as a hashtag cloud and a temporal rendering of tweet behavior patterns (MoKumax 2015).
- Buzzsumo offers Twitter influence metrics that include an account's number of followers, retweet ratio, reply ratio, and average retweets (buzzsumo.com). Buzzsumo also rates accounts based on Page and Domain authority. Page authority is an algorithm used to determine how well an item will be parsed by a Google

search. Domain authority applies this same algorithm to an entire domain. Buzzsumo also allows you to filter entity types, including influencers, bloggers, commercial entities, journalists, and so-called regular people. Buzzsumo also enables the user to toggle off broadcasters, by screening accounts with a reply to tweet ratio below 4%.

Other sites such as Bluenod (bluenod.com) or Mentionmapp (mentionmapp.com) advertise an ability to find influencers by constructing a social graph from interactions between accounts and shared use of hashtags.

2.3.2. Server-Side Social Media Analysis Tools

Social media analysis tools differ from the social media metrics previously addressed, insomuch as they offer analytic functionality to the user instead of the input-output interface more characteristic of the previous class. Social Media Analysis Tools are also often offered as a subscription service that allows the analyst to examine individual social media events. These tools are also not entirely focused on the social network and influence analysis of social media—adding trend, geospatial, and content analyses. While not an exhaustive list, the following applications are representative of contemporary market offerings:

- SnapTrends bills itself as platform capable of performing geospatial analysis on manifold social media sources. Their grouping function allows you to use cues from social media metadata to identify influencers present at a location or associated with a key term (snaptrends.com). A prime selling point of this service is its ability to use inferred spatial insights, beyond those mined from unstructured text. This includes locations derived from social network analysis and ambient mobile communications infrastructure, namely cell towers (Burris 2013; Tucker 2014).

- Babel Street, enables analysis of manifold standard media and social media sources across a significant portion of the language spectrum, with functionality that includes a geospatial component, sentiment analysis, as well as user-defined filters and cues that can be configured to trigger alerts. Babel Street also uses geo-inferencing to locate social media users who are not geo-tagging their content (babelstreet.com).

2.3.3. Social Network Analysis Software and Practical Application

As client-side software, this class is not inextricably linked to social media as a source of social network information, and can ingest various forms of social network data. Among these data types are association tables, which were used in this study to transfer content from the Twitter API into an analytic environment. The mechanics of this transfer are covered in the analytic methodology chapter.

Principal social network analysis was performed using the Organizational Risk Analyzer (ORA), part of the Carnegie Mellon CASOS tools (Carly et al. 2013). However, while base-layer social network analysis metadata creation was conducted using ORA, functionality introduced by this study does not exist as a module in ORA, or other contemporary social network analysis software.

ORA does offer extensive GIS capability when compared to its social network analysis software rivals. Like other such offerings, however, much of this consists of enhancements to the analysis of interplay between existing nodes and edges—social network analysis that is spatially clad, not necessarily a bottom-up GIS. ORA offers the user the option of displaying the network on a map if spatial metadata exists, aggregating by locations per user proximity specifications, and the ability to perform spatiotemporal analysis on this content using the "Loom" extension that displays movement paths, referred to as trails, between entities, and the locations to which entities have been associated in some fashion.

Regarding the specialized options most applicable to this study, ORA provides a geospatial assessment option that computes the distance between entities and applies this to existing connections. There is also a "Detect Spatial Patterns" report that accounts for clustering of nodes based on similar attributes, and a Twitter option that uses metadata to identify key accounts, tweets, and associated areas of influence. ORA also offers spatial versions of other existing network centrality measures. These options include:

- Closeness Centrality, Spatial – the distance from a node to edge-adjacent nodes
- Betweenness Centrality, Spatial – a computation of how many spatially distant nodes a particular node connects, presuming that such nodes have a penchant for influence or leadership
- Eigenvector Centrality, Spatial – score increases based not only on a node's centrality, but also on the centrality of its neighbors. The spatial version of this measurement determines the sum of eigenvector scores at a particular place, and scores those nodes accordingly

Other social network analysis tools examined include Palantir, which offers GIS functionality and a social media-specific extension known as Torch (Maher 2014), Dark Web (Chen 2011), and Analysts Notebook. In a review of available network analysis tools capable of conducting Dynamic Network Analysis (DNA), which differs from SNA in incorporating all elements of a network, and not merely the nature of relationships (Brieger et al., 2003), Maersk McKiney Moller Institute enumerated the benefits of social network analysis platforms, and suggested changes based on noted deficiencies (Wiil 2013). Among the improvements broached was an enhanced predictive analysis capacity, only observed within CrimeFighter Investigator (Petersen 2013) as it pertains to detection of missing links (Rhodes 2011). Beyond solely a statistical approach, this study offers prediction in the form of spatially-derived network development, and would be a novel addition to third-generation social network tools aspiring to meet predictive goals set forth by the Wiil study.

2.4 Network Latent Variable and Predictive Models

The concept of a dark network entails a series of relationships that is directly hidden or obfuscated as a result of some deliberate network impetus. This is not an inherently nefarious practice. However where it is in the case of terrorist or criminal networks, methods used to identify unknown social network variables are key to assessing a group operational capacity.

Examining unknown social network variables, understanding and quantifying the impact of physical distance on social interactions is key to social network studies such as social link prediction and social tie strength inference (Butts 2003, Daraganova et al. 2012). This inference is essentially reducing network uncertainty, both socially and geospatially, through the analysis of incomplete information. Numerous models of this type are addressed below, and are illustrative of various means of gaining analytic insight through the melding of social network and geospatial analysis.

The outlined ESSLVM is intended to identify unknown or underexplored social connections based on geospatial variables, and as such the preponderance of attention is placed here on similar models. Other systems however, use a similar interdisciplinary approach to geolocate points or entities that are otherwise unlocated via their social connections. This is referred to as spatial inferencing, which is the derivation of geospatial content from ambient data or metadata (Cheng et al. 2010). This section includes a review of both such latent variable models to underscore how complimentary both social network and geospatial analytic methods are.

2.4.1. Spatial Inferencing: Resolving Unknown Locations

With only a fraction of all social media events containing geospatial metadata, or geo-tags, methods have been developed to track nodes and events that are unlocated using spatial inferencing. Network or graph-based spatial inferencing methods differ from content assessments insomuch as graph assessments extrapolate information from a user's network to fix a position, while content-based work focuses on unstructured text, "I'm at the ballpark," "I'm at the beach," and so forth.

Other forms of spatial inferencing could include landmark colocation, popularly referred to as check-ins. Leveraging landmarks to conduct spatial inferencing along with a graph-based approach has also improved upon content-centric assessments (Yamaguchi et al. 2013). As an example of the effectiveness of a network-centric approach, using provided addresses of facebook users and their known friends, locations of users with undisclosed addresses were approximated with greater accuracy than could be achieved via IP address (Backstrom et al. 2010).

Another model (Cheng et al. 2010) analyzes a network of Twitter user mentions to zero in on an unlocated user's location. Using this approach, as of its publication, the Cheng model boasted creation of the largest and most accurate collection of gelocated tweets—110,846,236 Twitter users at city-level.

Further underscoring the significance of social network analysis in GIS, a different latent variable model type concerns the prediction of an event, its location, and its participants, by modeling the spatiotemporal dynamics of social networks (Cho et al. 2013).

This is made possible by inferring key information from pairs of actors and their involvement in events per observations that the spatiotemporal patterns of these relationships are not statistically independent. In other words, as a self-exciting process, the Cho model is contingent on temporal dependencies wherein the occurrence of one dyadic event precipitates another or numerous others. This model's claimed applications range from proactive policing, to the analysis of disease propagation (Cho et al. 2013).

An alternative approach (Li et al. 2013) that shares lineage with the Cho model, also focuses on linking events via shared actors to address latent event variables, but adds to spatial inferencing, network parameter inference—or in ESSLVM parlance, the reduction of network uncertainty. The Li study uses an armed conflict vignette to assess the causal interconnectivity of hostile engagements between US forces, the Afghan Government, Civilians, and the Taliban.

Spatial inferencing models emphasize that significant insight can be achieved melding temporal, social, and spatial methodologies, and that the underrepresentation of any one component can be compensated for by careful analysis of the others.

2.4.2. Network Prediction

With network analysis as the backbone of influencer identification, the more developed a social network is, the more accurate an influence assessment stands to be. Network prediction can rely on non-spatialized network analysis by modeling random progressions through the network, referred to as random walks, which are most likely to lead to new relationships as a result of edge attributes (Backstrom, et al. 2010). This and other supervised and unsupervised processes (Davis et al. 2013) can be effective, but are improved upon with the introduction of spatial components.

2.4.3. Optimal Network Prediction Using Spatial Content

Of the latent variable models examined, those most relevant to this study are the geospatial network prediction models. These models ascribe significant value to the distance between nodes in physical space and the likelihood of virtual social connections, whether they be latent or potential. Using this modality, the spatial distance between nodes in a network has been demonstrated as a viable means of reducing network uncertainty (Brieger et al. 2003).

FLAP (Friendship + Location Analysis and Prediction) uses a multimodal approach to identify social ties as latent variables (Sadilek et al. 2012). Testing this model on Twitter datasets, the different FLAP network uncertainty reduction modalities include leveraging knowledge user's location, their reciprocal connection patterns, as well as message content analysis as a function of belief propagation—or diffusion. To substantiate the accuracy of Flap predictions, assessed connections were scored against known hold-back connections. When starting with Twitter node-sets for Los Angeles and New York, ranging from edgeless (no connections) to 50% of all edges, FLAP's aggregated approach outclassed comparable models tested (such as Crandall et al. 2012; Taskar et al. 2003) by a significant margin—reconstructing a social diagram where only 50% of edges are provided. In this example, as a test of the model's ability to identify true positive edges in comparison to false positives, FLAP was .95 below the receiver operating characteristic (ROC) curve. This was accomplished with a precision-recall breakeven point of .65, meaning that 65% of unknown edges were identified without a false positive. When focused on a specific clique, or community, as opposed to the complete dataset, FLAP results were even more precise.

Sadilek et al's comparison of FLAP to the Tandar model (Tandar et al. 2003) showed that Tandar did not scale to the same extent as FLAP, and it requires some initial edge input at all levels. Comparison to the Crandall model demonstrated the triumph of FLAP's socio-spatial approach over a spatially dominant method, leading to the authors' conclusion that proximity alone is inadequate for the task of social network edge prediction--citing community space routinely shared with strangers as a prime example (Sadilek, et al. 2012). Emphasizing the reciprocal benefits of social and geospatial analysis as well as the methodological interconnectivity of all aforementioned latent variable models, in addition to excelling at social network prediction, FLAP was 84% accurate at plotting the location of unknown users, when an unlocated user has nine geolocated friends.

2.4 The Ethics of Going Socio-Spatial

The use of bulk, public, social media datasets as outlined above allows researchers access to vast amounts of data. That data however, is representative of people, and assessments derived from their aggregated content may offer insight that individuals never intended to divulge (Kadushin 2012). Additionally, the analysis of latent variables renders perceptible information that users may have deliberately kept private—such as their locations or relationships (Sadilek, et al. 2012). While the ability to defeat security controls has benefits in the realm of counterterrorism, efforts to safeguard a standard user's social media privacy (see for example, Weidemann 2013) are also susceptible to this technology.

According to the Federal Trade Commission, where a social media service furnishes metadata, and that content is handled in a manner keeping with the service's terms of use agreement, users that have agreed to the privacy policy but are otherwise averse to the manipulation of their personal content have no legitimate grievance (Claypoole 2014). Social media analysis service SnapTrends spatially renders social media that is enhanced by location inferencing, and in support of this practice the company has staked out defensible legal ground—asserting that users voluntarily post to a public domain and as such have no expectation of privacy (Sullivan 2014). Using similar reasoning, the study described here, does not entail direct interaction with subjects and uses only publically available data, and has been deemed exempt from further IRB review.

2.5 Summary

Given the large body of research on social network, a selection of which is cited above, it is important to note that the ESSLVM outlined in this study is principally intended to reduce network uncertainty and identify key influencers. This is demonstrated by constraining social networks using human mobility parameters that serve to physically delineate social communities, and thus their most significant actors. Literature on community detection and human mobility assessments addressed heretofore informed this method, however analytically ESSLVM is not intended to advance such processes, aside from demonstrating novel practical application of both. The next chapter outlines the means by which ESSLVM applies social network and latent variable concepts in a novel manner.

CHAPTER 3: ESSLVM ANALYTIC METHODOLOGY

The overall intent of ESSLVM is the advancement of propinquity effect as a means of solving latent variable problems. There are many components in this methodology, which are described in detail in Chapter 3 and 4. However, the finer points of ESSLVM are introduced by outlining the design principles which guided the development of the methodology.

3.1. Principles Guiding the Design of ESSLVM

ESSLVM methodology was intended to contribute to social network and geospatial analysis at multiple levels. This was accomplished by understanding the interplay between an individual's position in a social network and their relationship with others in physical space. These geospatial relationships contribute to an understanding of social network influence and the inference of social connections.

In the context of material addressed in the previous chapter, likely contributions include socio-spatial concepts, social network analysis applications, and spatial latent variable models.

3.1.1. Contributions to Socio-Spatial Concepts

By exploring socio-spatial concepts, ESSLVM is intended to demonstrate quantifiably that social network dynamics have a reciprocal relationship with physical distance vis-à-vis individual entities interacting with one another on a geospatial plane. This is in effect an advancement of the concept of propinquity effect, wherein new social network information is synthesized from corresponding geospatial content. This approach uses geotagged social media to make social network inferences at a highly granular scale of analysis.

3.1.2. Contributions to Socio Media and Social Network Analysis Applications

While many analytic modalities outlined in this study have been automated as tools using the Python language, ESSLVM is not intended to directly compete with social network analysis applications. Instead, ESSLVM propounds functional concepts that could be adopted and integrated by existing social media metrics—namely online services that score individual social media accounts. Additionally, ESSLVM is intended to augment known social network analysis modules aimed at identifying key influencers and inferring unknown social network relationships.

Many social network analysis processes in existence are constrained by the social network and its edges, and as a result are not completely spatially aware—though the ORA process of binding nodes by location, and offering a geospatial assessment at least approaches this concept. ESSLVM develops connections independent of known network adjacencies that can be subsequently reintegrated into the social diagram from whence they were derived.

3.1.2. Contributions to Latent Variable Models

As a latent variable model, ESSLVM underscores spheres of influence, and the individuals that comprise them. There are numerous highly advanced latent variable models in existence. While not yet quantifiably commensurable with these models, ESSLVM pilots concepts that could be incorporated into a system of comparable methodologies.

Calling attention to differences between cited models and ESSLVM, the Cheng geolocation model (Cheng et al. 2010) used only Twitter user mentions to develop its network—citing issues with the Twitter friends and followers API. User mentions constitute perhaps the highest quality node information. From a relationship standpoint, each user mention is a very contemporary act of volition by the account holder and not a relic from the social dustbin. This content however, is not always adequate in more sparsely populated areas. It is also not appropriate for instances where a network is to be developed from a seed account bereft any user mentions, such as the official media apparatus of a terrorist network. Because of their abundance in the Ferguson, Missouri vignette, however, user mentions were leveraged as a means of testing ESSLVM spatial association efficacy.

Alternatively, the Sadilek model (Sadilek et al. 2012), uses only friends, which is preferred from a qualitative standpoint, but for the same reasons listed above regarding data-sparse environments and non-interactive, high popularity users; ESSLVM expands on this approach by using both the friends' and followers' networks to ensure the social network diagram is as robust as possible. As an example of the utility of this approach, the followers' network would best bind two geo-tagged accounts that are following a third account that serves as an official media outlet and propagates extremist content, but that is not geospatially collocated. This places more of a burden on geospatial criteria within the model to define a community and identify potential relationships.

Additionally, in the Sadilek model, it is assumed that a user remains at a location until they tweet again from a new location. However, in trials conducted where data is not as plentiful as Los Angeles or New York—the test bed of the Sadilek model—there must be controls in place that will allow the user to account for pronounced asynchronous presence where a paired event might occur days or weeks after an initial event.

Finally, a key analytic undertaking was the development of social network discovery methods likely to be conducive to influencer identification and supported by statistical outputs indicating focal point-centric clustering of social network centrality scores in space. While its output cannot be directly compared with latent variable models cited in the previous chapter, ESSLVM use a social connection prediction approach that is conceptually comparable. Examined latent variable models however, do not employ methods that take into account the effects of attraction and repulsion on entities distributed in physical space per their social network variables, which is an underpinning of this study.

3.2. Overview of Methodology

The many interrelated components of ESSLVM are depicted in Figure 4 and details of their functions are explained in the following sections.

The analysis of social media can often be subjected to two major logistical obstacles: Procuring the data and handling it at scale. Thus, a portion of this methodological overview details necessary preliminary steps, including use of additional supporting Python scripts.

This data is in turn geospatially correlated to an area of interest in order to identify a community. From this community, specific entities (i.e. Twitter users) are identified as seed accounts—correlated to the area of interest—and the Twitter Friends and Followers API is queried in order to develop a social network from these seeds. Using this network, the connections of entities as nodes in the social diagram are constrained to other entities within defined spatial extent, in order to underscore local relevant network clusters and associated influencers.

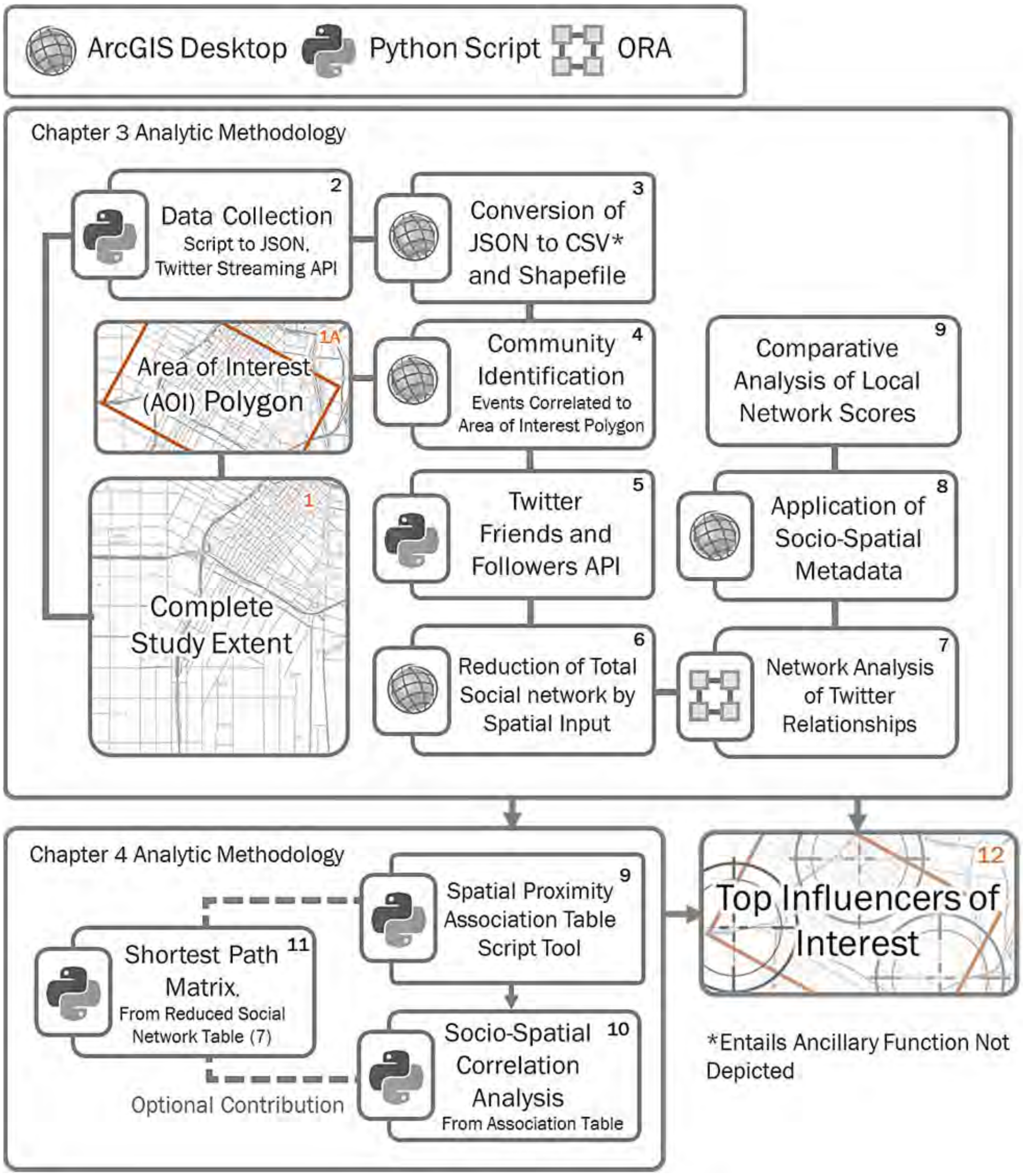

Figure 4 ESSLVM Conceptual Model, Analytic Methodology to End-State

It is from this process that we arrive at a means of substantiating and validating the hypothesis that position in a social network has bearing on an individual's relationship with others in physical space, and as a result, individuals or organizations postured to influence a network via direct conduits such as local leadership figures and on-site organizers, possess a qualitative advantage. Additionally, using the spatially biased social network as an output, more advanced methodologies addressed in Chapter 4 can be applied.

3.3. Initial Data Collection

Twitter has authorized select vendors to sell this data, however prices for a one-time provision of geotagged data bundles are likely impractical for those not engaged in manipulation of the content for commercial ends, or who are not otherwise externally funded. For example, as of late 2014, one such social media content vendor, Datasift, offered geotagged tweets at an out of contract minimum of 1,000 US dollars, covering 40 consecutive days and up to 1,000,000 total tweets (datasift.com).

This study's Twitter data requirements were satisfied using a Python script developed for this project to access the Twitter Streaming API (Twitter 2015), which is a public source of information that is free of charge. By virtue of its streaming output, retrospective queries are not an option, which necessitates anticipatory collection on events of interest. Data was extracted in JavaScript Object Notation (JSON) format to a text file, and consisted of geotagged tweets within a user-defined bounding box. For this project this script was organically adapted to function as an executable file, requiring the name of the new output JSON and coordinates for the southwest and northeast corners of the bounding box. The script continued running until manually discontinued, and while running, the JSON was populated in near real-time. As an ancillary process within the script, the Twitter Streaming API JSON was converted into a comma-delimited table, cleaned of spurious bot-generated tweets, and converted into vector data that was subsequently manipulated in ArcGIS Desktop.

As an example used to illustrate the various steps of this methodological approach, Twitter data was pulled for the 13 November 2014 USC vs. Cal college football game for administrative parcels immediately adjacent to the USC campus and Los Angeles Memorial Coliseum—though when using the Twitter Streaming API there tends to be significant spillover outside the specified bounding box. This Twitter steaming API pull occurred between 16:00, 13 November and 01:40, 14 November, resulting in 4414 total geotagged tweets. These tweets are depicted in Figure 5 below.

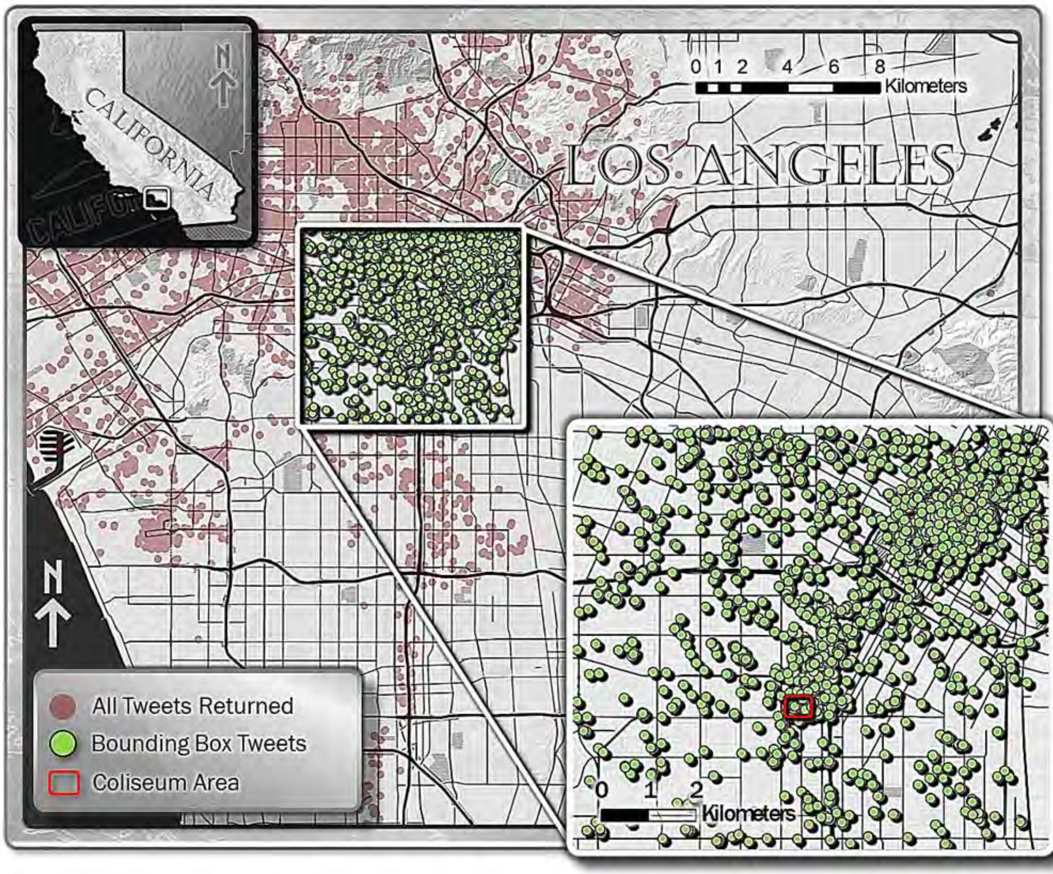

Figure 5 Tweets Collected in the vicinity of the Los Angeles Memorial Coliseum Collected during a USC Football Game

3.4. Community Identification Using Geospatial Correlation Methods

From within the total dataset pulled from the Twitter stream, correlating spatial subsets of the collected vector data to a particular smaller area of interest enables the distilling of content into a community that can be analyzed. It is important to note that the study extent, used to pull the tweets, can be dictated by the area of interest polygon, or vice versa depending on the whether the area of interest polygon cued interest in the space around it, or if you are looking for a defined area of interest, unknown at the outset, inside a known study extent.

Since this smaller area is considered discrete state space, this method requires discrete state space membership of Twitter point data, e.g. points intersecting a polygon, such as a specific building whose patrons are the subjects of study. Distinct social events can also be distilled with the addition of discrete temporal limits, essentially bracketing the event within a time slice and demarcating its footprint using a polygon.

Once all events within the smaller area of interest have been identified, the process then determines all the unique social media accounts within that polygon, and interrogates the complete extent of the data for the same account information. Using the example of tweets collected for the USC vs CA football game, the polygon was defined by the football stadium, Los Angeles Memorial Coliseum, only taking into account the temporal span when the venue was open in association with the game. Using this extraction method, there were 275 total geotagged social media events generated by 175 unique Twitter accounts active at the game. These accounts were then correlated to all tweets collected within the USC campus bounding box, amounting to 383 total social events generated by accounts that were spatially and temporally correlated to the football game. The distribution of these points is shown in Figure 6. If the social media events had been distilled across the entire city, this process would identify locations around the venue where fans were before and after the game, e.g. tailgating locations, residences, etc.

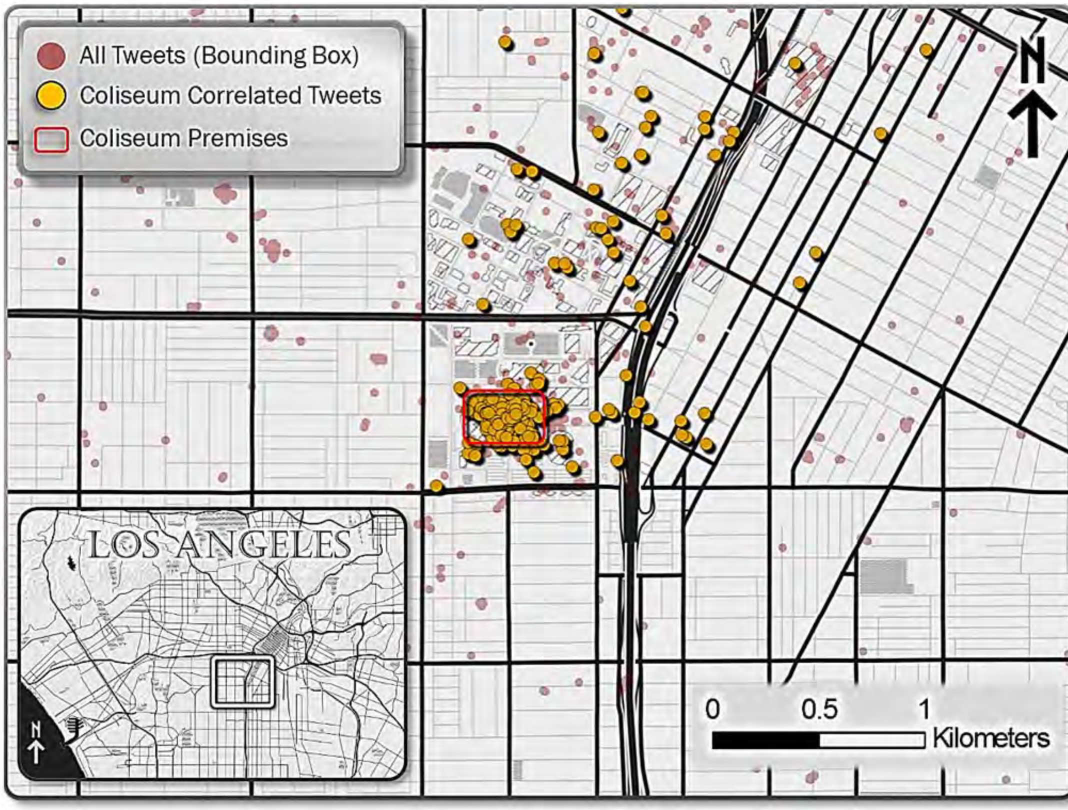

Figure 6 Tweets Correlated to the Los Angeles Memorial Coliseum

Another method of depicting the number of different users occupying the same space is by binning. Hexagon binning to visually render Twitter entity diversity spatially is depicted in Figure 7.

Figure 7 Tweets Binned to Depict Entity Diversity (Number of Entities per Bin)

From the tool output, you can infer answers to questions such as: "Where do accounts that have been active on campus live, dine, and engage in recreational activities?" Or, in a different context, "what accounts are active on both sides of a national boundary, and where do these accounts converge?" Documenting this type of activity over time can be key to developing patterns of life of select individuals or groups.

3.4. Creation of a Spatially Reduced Association Table and a Socially Enhanced Vector File

The output from the community definition process provides a list of accounts associated with the area of interest—in this example it was 175 unique Twitter accounts active within Los Angeles Memorial Coliseum during a football game. These accounts serve as seed accounts—or the accounts used to query the Twitter Friends and Followers API.

The Twitter Friends and Followers API (provided as open source by Twitter) allows access to a database of both reciprocal (friends), and unilateral connections (followers). However, it is limited in the volume that can be accessed at any one time. In this application, a Python script was used to access the Twitter Friends and Followers API, overcoming API imposed volume restrictions with built-in cursor functionality that "sleeps" when a data ceiling is reached, and resumes pulling data after the requisite waiting period has elapsed.

The Twitter Friends and Followers API output is a JSON for every seed Twitter account. These JSONs are then converted into lists of all friends and followers for each of the original seed accounts. An additional Python script appends the associated seed account name to each row of their friends and followers list, and joins all disparate lists together to produce an association table. For example, if Account A has 25 followers, their Twitter friends and followers are written to a text file, and converted to a table with 25 rows each containing one friend or follower's unique account information or Twitter "user screen name," and the original account, "User A" or seed account. An example of this association table output is depicted as Table 2.

Table 2 Association Table Example

Screen_Name	ID_str	Statuses_count	Followers	Friends	Seed
Screen_Name_21	XXXXXX02	1056	5843	6175	Screen_Name_1
Screen_Name_22	XXXXXX31	49319	11393	248	Screen_Name_1
Screen_Name_23	XXXXXX71	13981	683	140	Screen_Name_1
Screen_Name_24	XXXXXX18	107	246	800	Screen_Name_1
Screen_Name_25	XXXXXX28	61	63	254	Screen_Name_1

The association table can be ingested into social networking applications. This process was repeated for every account active in the football venue polygon, and all disparate friends and follower lists were aggregated into a master association table. The result from the football game example was 46,030 friend and follower relationships associated with the original 175 seed Twitter accounts—totaling 37,353 unique Twitter accounts.

As an association table the data can be imported into the ORA social network analysis application. When an association table containing pairs of related entities is imported into ORA, each row becomes a dyad, two nodes connected by an edge. From this list of dyads, a multi-tiered social network diagram is formed from the original seed account relationships represented on the association table. However, while each seed account is selected based on geospatial correlation to the polygon of interest, their friends and followers may well span the globe. As a result, this step can often be functionally impractical until the network is reduced per spatial criteria in subsequent steps. This is because of the processing power and memory required to turn hundreds of thousands of relationships into a comprehensive social network diagram.

3.4.1. Development of Socio-Spatial Metadata

When the objective of a study is the identification of locally relevant influencers, widely distributed networks can be culled by removing nodes beyond a distance that defines a relevant "mobility space." For the football game example, the intent was to reduce the 37,353 unique entities on the association table, to those entities who are actually active in and around the study extent—adjacent to Los Angeles Memorial Coliseum.

How can this distance be defined? The exhaustive study of human mobility space is an undertaking worthy of its own thesis, and was not a research objective of this one. However, there are many studies which provide good estimates of appropriate distances. As an example, for the vignette described in the following chapter, the results of Becker et al.'s study on cellular telephone derived human mobility characteristics provided a basis for selecting a mobility limit for daily median weekday range in an urban area of approximately 5600 meters (Becker et al. 2013). To confirm this is an appropriate generalization, standard deviational ellipses for all events produced by each of the seed accounts can be computed and compared to the Becker-derived, or any other pre-determined mobility buffer. The spatial distribution of each account's events should fall approximately within the buffer. For the football game example, the original bounding box was used as a surrogate for mobility space.

3.4.2. Isolation of the Spatially Relevant Network from the Complete Social Network

Next, all social media events collected throughout the entire mobility buffer are correlated to the list of friends and followers associated with the original seed accounts. In joining a table of entities inside the mobility buffer to the master association table, only matches are retained. This results in the identification of secondary connections—accounts not present in the area of interest polygon—that still share a social connection with seed accounts that were correlated to the area of interest polygon. Thus you have a new set of users who did not tweet from the initial polygon of interest, but who were socially connected to the seed accounts inside the polygon of interest, and whose geotagged tweets have indicated they were located in the mobility buffer.

This method achieves a significant reduction in overall network size by restricting relationships that are likely most relevant to the initial polygon of interest. In the LA example, using only Twitter friends (reciprocal ties), those seed accounts likely attending the game had only 149 social connections inside the initial bounding box. The resulting streamlined association table is referred to here as a spatially reduced association table. Figure 8 is a depiction of these nodes in a social graph. An output of this process is a shapefile that includes all social media events generated by these entities, inside the mobility buffer.

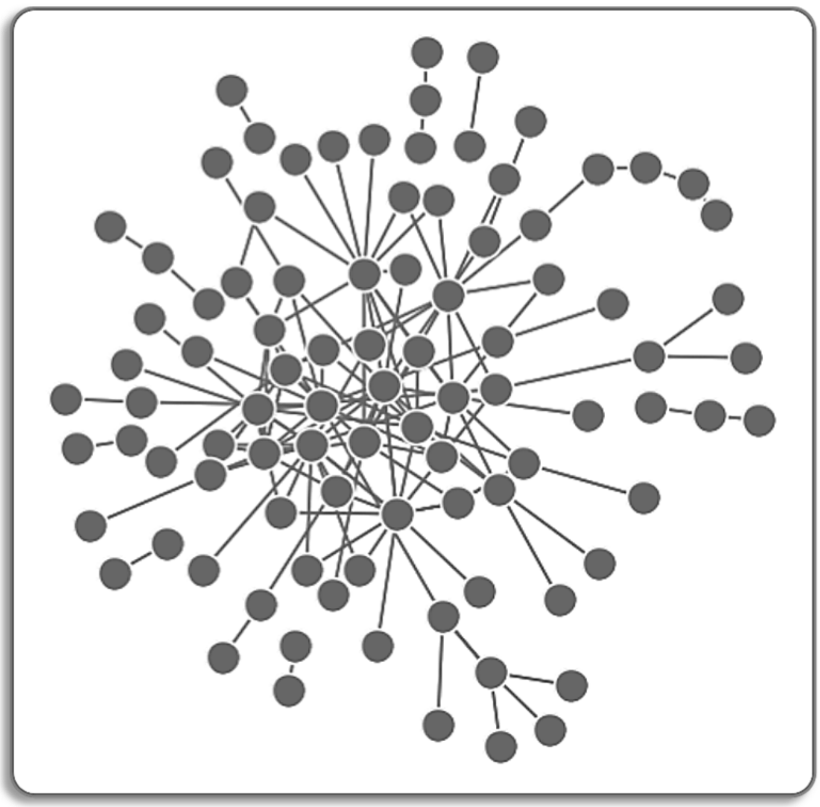

Figure 8 Social Network Graph Depicting Twitter Friends of those Attending the USC-CAL Football Game

3.4. Development of Socio-Spatial Metadata

Once in ORA, the social network diagram produced from the spatially reduced association table can be analyzed for various measures of network centrality. The intended benefit of network reduction through application of geospatial constraints is to bias these social network scores in favor of entities with more local connections than distant ones, so that locally relevant influencers might be more apparent.

Derived from the spatially reduced association table, the complete gamut of social network centrality measures produced by ORA—which amounts to 39 in total— are exported as a table. The social network centrality scores for each analyzed Twitter account are then joined to social media event shapefile attribute data, to create a social network analysis enhanced vector file. With all Twitter accounts identified in previous steps now joined with their social network centrality scores as metadata, spatial visualization and analysis of social network values associated with the event vector layer can occur. As an example in Figure 9, the nodes associated with the USC vs. CAL game (in Figure 8) have been scaled by their betweenness centrality. As spatial metadata, such metrics and symbology can be appended to the resulting shapefile, as depicted in Figure 10, where points are also symbolized by their betweenness centrality. Of note, in Figure 10, there are additional entities visible which were not correlated to the football game, but have reciprocal social ties to entities who did.

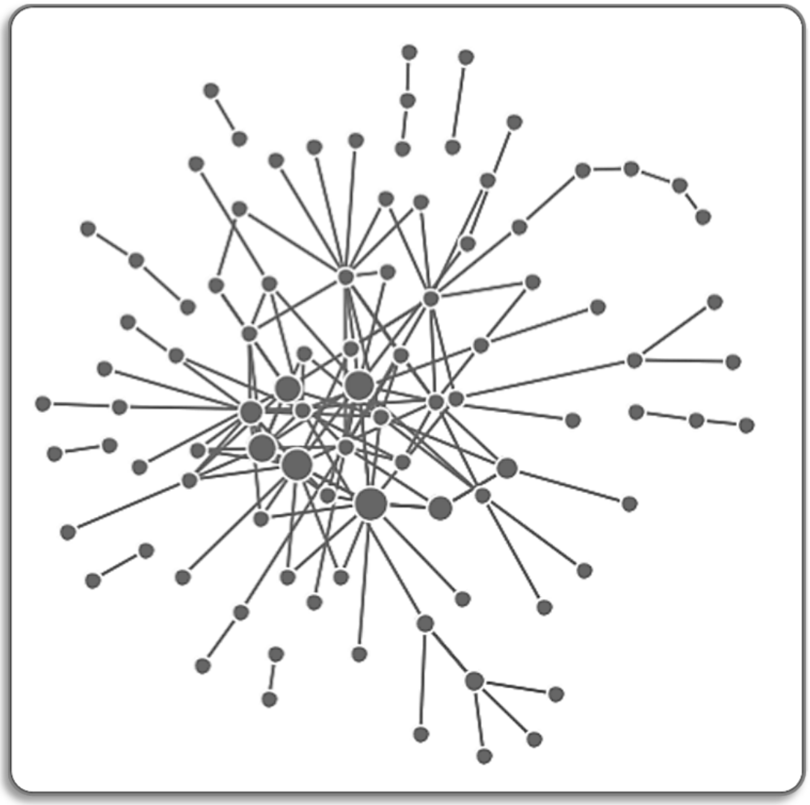

Figure 9 Coliseum Correlated Nodes Scaled by their Betweenness Centrality

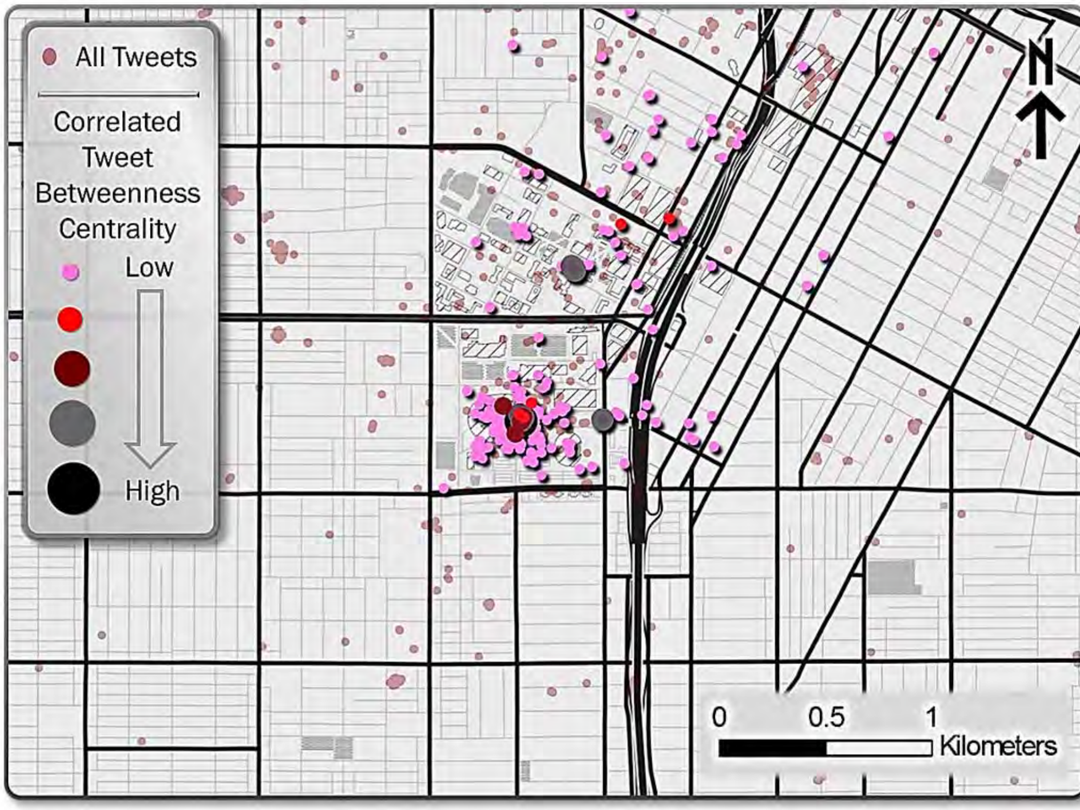

Figure 10 Coliseum Correlated Social Media Events and Reciprocal Connections, Scaled by Betweenness Centrality

The pairing of social network values to Twitter event vector points is also key to subsequent steps addressed in Chapter 4 that include the identification of socio-spatial relationships and analysis of the correlation between Twitter account social network metrics and their geospatial distribution.

3.4. Comparative Analysis of Twitter Accounts

Using spatially biased social network data to identify locally relevant influencers, additional analysis can include comparing other account content such as publically posted tweets and photos to network scores. This is simplified by sorting the range of network centrality scores to identify top scoring accounts in each category. In the case of the protest vignette described in the following chapter, top scoring accounts were researched to assess their role in a protest. As will be seen in Chapter 5, the process of subjectively analyzing entities vis-à-vis their centrality measures, was used to determine which centrality measures are most appropriate to identify entities actively influencing the protest. This comparative analysis however, was not applied to the Chapter 3 football game example.

If the key hypothesis of this research holds true, local influencers actively involved in their community generate higher social network scores within a social network parsed spatially—in the steps mentioned heretofore—than those who rely on a broader, more externally distributed support base.

3.4. Summary

The addressed stages of the ESSLVM process combined spatially correlated social media events with their non-spatial social connections to identify social relationships most relevant to a specified area of interest. From these spatially-biased social relationships, the identification of local influencers is assessed to be more likely. Applying tenets of propinquity effect and the mechanics of latent variable models, the data generated through this process can be assessed for other social insight per spatio-temporal dynamics. The implementation of this approach is covered in the next chapter.

CHAPTER 4: ADVANCED ANALYTICS AND THE ESSLVM PYTHON SUITE

Endemic Socio-Spatial Latent Variable Modeling (ESSLVM) Python scripts are designed to identify and quantify socio-spatial relationships from within the spatially reduced dataset addressed in the previous chapter. There are multiple Python scripts that comprise the ESSLVM script suite, with the keystone script an ArcGIS Script tool (under the prototype designation Blue Starling). Other associated Python scripts preform key data processing tasks but are not in an ArcGIS script tool GUI. The central component of all scripts is the creation and manipulation of an association table.

Beyond a development undertaking however, there is additional ESSLVM analytic methodology entailed in detecting relationships per user-defined social, spatial, and temporal criteria. These functions are inextricably linked to methodology outlined in the previous chapter—as they are intended to use the same social media output shapefile as a data source. Scripts outlined here perform additional analytic functions on this social media shapefile. To this end, ESSLVM is a process intended to enable follow-on social network analysis by underscoring inferred relationships. The scripts created for this study and their corresponding analytic underpinnings are as follows:

- *ESSLVM ArcGIS script tool (Blue Starling) association table creation.* This ArcGIS script tool uses a shapefile as input and identifies latent or incipient social connections based on spatial, social, and temporal dynamics—recording these inferred relationships to an association table as an output.
- *Socio-Spatial Correlation (Blue Starling) Analysis Functions.* This script uses the output of the previous script, and determines whether entities demonstrate socio-spatial correlation to one another using the distances between them, their social network centrality values, and Spearman's Rank Coefficient Correlation. The output

is an association table that includes their correlation coefficient and other supporting data.

- *Shortest path matrix.* This is a script that uses the NetworkX Python library to determine the distance between any two nodes in a social network. This method can be applied as a discriminator for other steps, e.g. keep only those nodes that are within 2 steps of one another in the social network. This script's input is the spatially reduced association table discussed in Chapter 3, and the output is a matrix that includes all possible node pairs.

4.1. ESSLVM ArcGIS Script Tool (Blue Starling) Spatial Association Table

While existing social network analysis software addresses the analysis of nodes and edges, this tool suggests new relationships based on specified spatial, social and temporal criteria—building an association table based on the addressed manifold relationships that exist between entities in physical space. Additionally, this tool includes another user-determined criterion—the number of encounters that must occur for a relationship to be considered significant. Considering shared public space, such as a coffee shop, at what point is there a reasonable expectation that two entities within that physical space share a social relationship and their meeting is not purely coincidence? These variables (spatial, social, and temporal) are parameters determined by the user. If these user-specified criteria are met, this process summarizes all these variables and adds them to a script- generated association table. This includes the average time difference between each of the dyadic node's sequential tweets, their average distance apart in physical space, and their difference in social network scores.

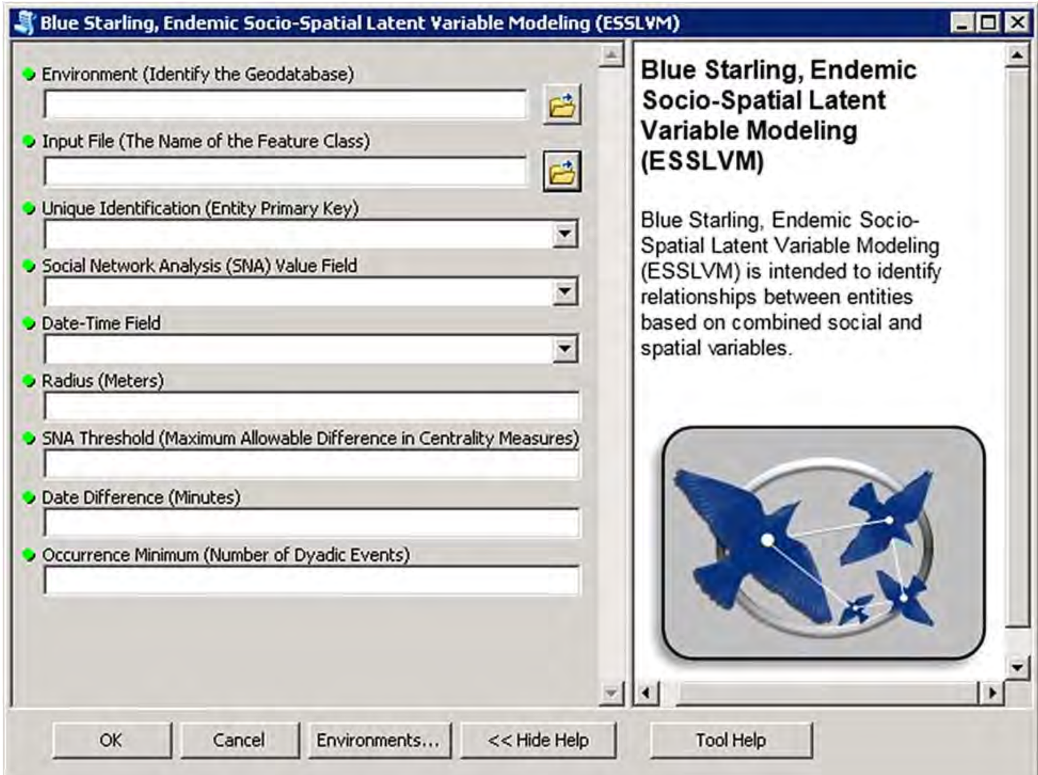

Figure 11 “Blue Starling” Script Tool

4.1.1. ESSLVM ArcGIS Script Tool (Blue Starling) Spatial Association Table Parameters

The field variables in the script allow the user to specify which of their fields meet these criteria, without having to manually alter the attribute table in advance. These include “ID” field, which is any field that uniquely identifies each account. For Twitter, that would be the “User ID” or the User Screen Name.” The “SNA” variable is the preferred social network centrality criterion for establishing social distance, and the “DateTime” field variable is an ArcGIS-recognizable field that contains both date and time information. Of note, the standard Twitter date-time format does require modification.

Other non-intrinsic variables serve as criteria that determine what inter-node dynamics will qualify nodes as members of a dyad, or to be retained as a pair to undergo the subsequent socio-spatial correlation analysis process. Using these variables, all entities are compared to one another to assess relationships, or the prospect thereof. These variables include:

"Radius," is the maximum geospatial distance within which two entities will be considered a dyad, it is the principal disqualifier, and no other calculations will be conducted if data does not meet this criteria.

"Closeness," closeness degree centrality, or another "SNA Threshold" which is the maximum allowable difference centrality between two nodes in a social network. The measurement used in the vignette was eponymous closeness centrality, as it is a measure of a node's mean social distance to other nodes in the social network, and as a result. A relatively large difference in closeness centrality is indicative of corresponding significant separation between two nodes in the social network—per each nodes average distance to all other nodes in the network. This does require some advanced knowledge of social network metrics, and the centrality variable is the user's prerogative. The process of comparing closeness centrality scores however, can be refined using a true measurement of the social network distance between two nodes as steps represented in a matrix. Such a matrix is representative of all shortest paths between every node in the social network, and is described in a subsequent chapter.

"Date_Diff" is the maximum allowable time in minutes between events, or in this case tweets, that is acceptable for two accounts to be considered a dyad by computing the absolute difference between date-time fields. Because of the aforementioned asynchronous nature of certain social media datasets, especially in data-sparse areas, this variable can be set to account for relatively loose tolerances. In urban areas, or periods of intense activity, it is advisable to get the "Date_Diff" variable as low as possible to optimize accuracy. This process replaced the discrete time slices of earlier diagnostic analysis to allow for assessment of an entire dataset in one run, and to negate time slice edge effects.

Variables that control the location of input and output data include arcpy.en.workspace or the geodatabase where the input file is located. This information is input by entering it in the “Environment” Field. It is recommended that each project be conducted in a dedicated geodatabase. The only allowable output is a feature class. “Input File” is the name of the shapefile that is being analyzed, and the output association table’s name is derived from this shapefile.

4.1.2. ESSLVM ArcGIS Script Tool (Blue Starling) Spatial Association Table Functions

There are several functions that occur once all the parameters are set and the tool is initiated. Using the provided shapefile, these functions create an association table, then scrub its contents per specified variables. An example of the output is depicted in Table 3.

Table 3 Spatial Association Table Output

SNA	Node_ID	SNA	Node_ID	FREQ	SNA_Diff	MEAN_Time	MEAN_Distance
0.00388	Account 1	0.00388	Account 2	54	0	13.175	9.654383
0.00388	Account 2	0.00388	Account 1	54	0	13.175	9.654383
0.00387	Account 3	0.00388	Account 7	66	0.00001	11.5	17.196278
0.00388	Account 4	0.00387	Account 5	66	0.00001	11.5	17.196278

4.1.2.1. Converting Fields

Using Esri geoprocessing functions from the “arcpy” Python library, the original file is copied to preserve its attribute data, and the copied file has the “SNA_Measure,” “Node_ID,” and “DateTime” fields altered from the user input attributes, to the field names reflected in the output using field management functions. This is a necessary step, to ensure script compatibility with a variety of attribute input names.

4.1.2.2. Point Distance Analysis

This function determines the distance between all points that fall within a user-specified radius of each other using point-distance functionality in ArcGIS. The radius is set in this script as a variable. In this model, the dataset is run against itself after it is copied. The output of this point-distance process is a table pairing every point to every other point. This table, the "spatial association table," is the destination of all output from the subsequent functions described below.

Since the output table consists only of the input point object ID, the near object ID and the distance between them, data from the shapefile attribute table is joined to this table twice, using the variable designated as the account-unique identification for both input and near fields. This allows all Twitter account information to be analyzed for relationships during subsequent processes. Additionally, the average distance for each dyad is calculated as "Mean_Distance" in Table 3. These average dyad distance values are added to the table redundantly for each row corresponding to the same dyad.

4.1.2.3. Relationship Metrics

To calculate relationship metrics, a new field is created in the spatial association table that contains a concatenation of the input and adjacent or "near" user screen names (the dyads). Because directionality cannot be assessed spatially, the same entities in reverse order are also counted and the amount of times each dyad in this field appears is summed using frequency analysis functionality. The results are joined back onto the main association table—in so doing, this creates a field that is assessed for removal of entity-homogenous dyads, or the same user screen name twice in the same field. This would occur when multiple tweets from the same account occur within the same time-span and specified radius. Only entity-heterogeneous rows are retained. This field is then used to perform a count of the number of times each of these accounts (as a dyad) meet other relationship specifications. The total number of times each entity meet association parameters is recorded as “Freq” for frequency, in Table 3.

4.1.2.4. Social Network Analysis Metrics

As a surrogate for the distance between two entities in a social network, the absolute difference between each account’s “SNA” variable is calculated, using Esri field calculations and added to the association table. Per the user-specified “Closeness” criteria, this absolute difference of SNA values is used determine which rows are retained. The difference between their SNA scores is recorded (as depicted in Table 3) as “SNA_Diff.”

4.1.2.5. Temporal Analysis

To determine the time between tweets that qualify a pair of entities as a social dyad, per the user-defined “Date_Diff” criteria, a time difference field is added, and the absolute difference between time fields is calculated to populate this new field using the ArcGIS date difference function. As is the case with the other criteria variables, only those rows that meet user requirements are retained. The average time difference between each dyadic pair is also calculated and added to the table redundantly for each row corresponding to the same dyad. As dyad refers to two nodes in a social network, all events generated by these still represent only one relationship. This is recorded on the output association table as “Mean_Time.”

4.1.2.6. Administrative Functions and Output

The final output for the spatial association table deletes all fields created as a result of multiple join functions, leaving only the input account, or entity, its SNA measure, the near account, its SNA measure, the number of times all criteria were met for each dyad, the mean distance between accounts in a dyad, the difference in social network values, and the mean time difference. As part of a clean-up function, all interim tables created as a result of this process are deleted, leaving only the input and two output files in the project geodatabase.

The socio-spatial analysis input table, essential to the next section, retains all measurements between accounts that meet criteria, with a row for each occurrence. This allows the near features' social network value as an X field, and the distance between accounts as a Y field, in the subsequent Spearman's Rank Correlation Coefficient process.

4.2. Socio-Spatial Correlation Analysis

Consider the crowd that gathers around a public speaker, a politician, an executive, or a celebrity—anyone that serves as an influencer or social focal point. Now take into account that influencer's social network values and the values of those they attract. These influencers, by virtue of their standing, are likely to have a higher social network score. Research that contributes to this study indicates that interaction between social entities in physical space can have observable and quantifiable patterns documented as spatial attraction or repulsion. These patterns can be evident in the signature of the influencer, or those consistently attracted to them.

Still a subset of Blue Starling, this script uses an association table derived from the previous scripts to determine the level of correlation between spatial distance values and social network centrality. It is assessed that this method constitutes an advancement of the community's understanding of propinquity effect and its bearing on groups of people vis-à-vis social network values. The intent is to detect concentrations of social influence on a geospatial plane so that these connections can contribute to traditional social network analysis.

4.2.1. Socio-Spatial Correlation Analysis Script Objectives

A now vestigial aim of this process was the geospatial analysis of crowd dynamics and macro-cognition (Huebner 2014), taking into account the tenets of swarm theory, wherein forms of incentivized attraction or repulsion can dictate ostensibly coordinated movements (Wang et al. 2013). These concepts however, were instead applied to the socio-spatial analysis addressed throughout this thesis—hence the name “Blue Starling,” coined after the swarm-like murmurations of Starling birds.

The principal objective of this analysis is to enable a user to take a shapefile enriched with social network metadata and determine whether or not the events that comprise this shapefile demonstrate significant socio-spatial correlation per Spearman’s Rank Correlation Coefficient. Spearman’s Rank Correlation Coefficient is a means of assessing the statistical dependence between two variables, denoted by its “rho.” This method is preferred over other similar approaches that use raw inputs as variables, because the ranking of values inherent to the Spearman process accounts for data that is not evenly distributed (Gauthier 2001). Spearman’s Rank Correlation Coefficient in its standard form can be represented using the following equation, where ρ = the rho, d_i = the difference between x and y ranks, and n = the sample size:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

During the development process, a prototype model separated an assessed influencer class from all other nodes using the social network values of the influencer class, and the distance of all other nodes from that class as input data for a correlation process.

This diagnostic model however, assumes knowledge of a top influencing tier per social network analysis, and requires knowledge of relevant time slices on which to conduct analysis. While the method was not retained, it did serves as a proof of concept, with the isolated influencer class demonstrating statistically significant correlation between their distances from ambient entities and social network centrality measures of those entities.

4.2.2. Spearman's Rank Correlation Coefficient Script Functions

This Python script is fed by a designed derivation of the Blue Starling script tool and as such is responsive to the same user-dictated criteria. To perform correlation analysis however, instead of summarizing each relationship—as is done in the previous step—the precise distance between each entity is recorded to the association table, and each such occurrence is retained.

Incorporating Spearman's Rank Correlation Coefficient into an automated process makes use of Pandas (pandas.pydata.org), a library of data analysis resources in the Python programming language, as well as SciPy (scipy.org), a library of scientific tools, also in Python. The script created for this study points to an input comma delimited table (CSV) created from the correlation coefficient input table that was an output of the previous step. Using read and "groupby" functions in pandas, the table is effectively split by a user designated unique identifier, in the case the input Twitter account, which is the user screen name, or node ID—per the lexicon of the previous script. This allows each of the ensuing SciPy functions to be run against each set of relationships independently and virtually simultaneously. The first of these is the "rankdata" function that assigns an ordinal value to each input value as a prerequisite for Spearman's Rank Correlation Coefficient. The next function is correlation, with a specified output of Spearman's rho, and the P-value, to quantify the statistical significance of the rho. As an output, all correlation values are appended to the original association table.

4.3. Shortest Path Matrix

Using the Blue Starling script tool, a surrogate for the position of a node in a network in relation to other nodes is a calculation of the absolute difference between closeness values—the average shortest path to all other nodes in the network. However, a superior means of assessing network proximity was developed using a script in conjunction with the NetworkX Python Library. This script creates a matrix based on all nodes in the social network and their shortest distance, in number of edges traversed, from one another. This matrix allows spatially derived relationships to be screened by the number of steps between them and other nodes, instead of an absolute different in centrality scores. This script can also be configured to only return a matrix with a specified minimum number of steps between nodes. An example output of a shortest path matrix is depicted using anonymized data, in Table 4.

Table 4 Shortest Path Matrix Example

	Screen_Name_1	Screen_Name_2	Screen_Name_3	Screen_Name_4	Screen_Name_5
Screen_Name_1	0				
Screen_Name_2		0	1	3	
Screen_Name_3		1	0	3	2
Screen_Name_4		3	3	0	
Screen_Name_5			2		0

4.4. Summary

ESSLVM advanced methodology is intended to leverage socio-spatial concepts to identify relationships where social connections are not known, or an appraisal of the significance of existing nodes is lacking. Ultimately however, the results are conceptual, and require comparison to traditional methods of social network analysis. This methodology is applied in the next chapter to identify key influencers and the entities that comprise their network, among the chaos of an active protest.

CHAPTER 5. VIGNETTE: SOCIO-SPATIAL EXAMINATION OF SOCIAL MEDIA AT THE FERGUSON PROTEST EVENTS

An optimal demonstration of ESSLVM is best found in an area where social media events occur in regular and frequent intervals. Such conditions can occur naturally in high traffic, bustling metropolises often used to demonstrate other latent variable models—New York or Los Angeles among them. These conditions however, can also be fomented by events, wherein the populace is whipped into a fervor and social media event frequency is intensified. The formation of crowds is characteristic of such a state. Enter the principal vignette for this thesis, the November 24th 2014 announcement of the grand jury decision on the Michael Brown case, the context of which is described below. Specifically of interest here was the crowd activity that occurred in and immediately around Ferguson, Missouri in association with this announcement.

5.1. Vignette Atmospherics, Setting the Scene

This study takes place in Ferguson, Missouri and adjacent townships—part of the greater Saint Louis metropolitan area. Data from Esri's Tapestry system was used to profile the study extent's socio-economic attributes (Esri 2015). Tapestry is a collection of geo-demographic data derived from different sources, including US Census data, the American Community Survey, Experian's INSOURCE consumer database, and other consumer surveys. These data are used to extrapolate lifestyle segments that serve as a composite characterization of all input data. The extent of this study included three principal Tapestry segments: Family Foundations (21%), Metro City Edge (19%), and Modest Income Homes (17%). Household Income for these segments range from approximately $20,000 to $38,000. Common attributes associated with these segments are predominately black populations with socioeconomic or demographic issues that are pronounced by national standards. These issues can include single parent homes, low median education, or an above average number of recipients of public assistance.

5.1.2. Ferguson, Missouri Events Resulting in Protests and Subsequent Riots

While the focus of this study is not the sociological implications of events surrounding the death of Michael Brown, the selected vignette does address crowd behaviors precipitated by this act, and the documentation of associated events does provide valuable context. The temporal span of the study includes the evening of the 23rd and into the morning of the 25th of November, 2014. On the evening of 24 November, a grand jury delivered its decision on the alleged wrongful death of Michael Brown, a resident of Ferguson, Missouri, who was it was determined, justifiably killed while resisting arrest on 9 August 2014. The August event resulted in national outrage and allegations of institutionalized racism, likely as a result of the perception that a white law enforcement official responded with disproportionate force while subduing a black teen—Michael Brown. Though the issue is more nuanced, this thesis is in no way intended to assess culpability, and only uses crowd events associated with protests as fodder for a socio-spatial study.

In the days leading up to the grand jury's decision, protests in front of the Ferguson Police Department intensified. Protest activity in this area is depicted in Figure 12, which depicts protest Tweets by the variety of different users using point statistics.

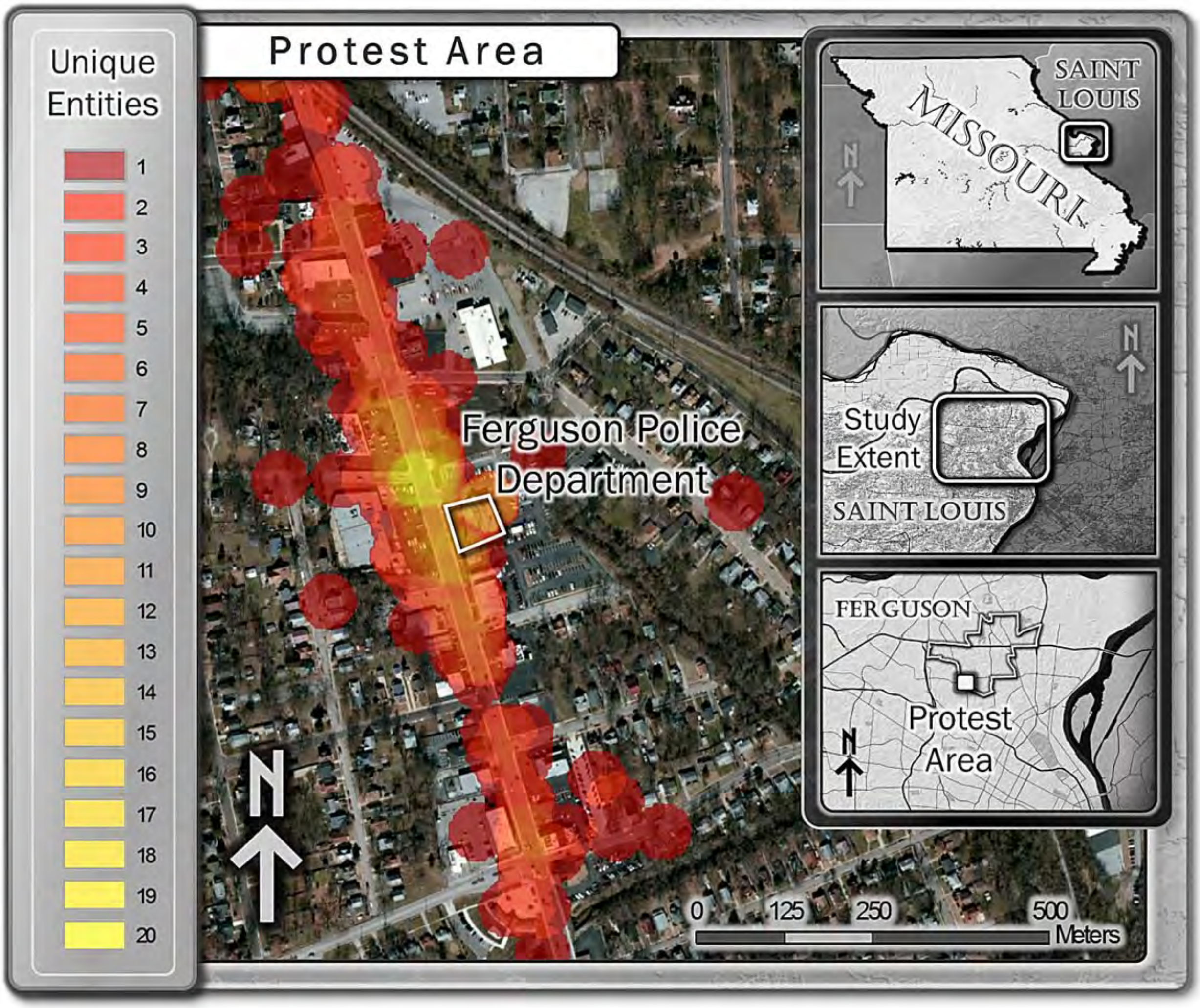

Figure 12 "Twitter Activity in the Ferguson Protest Area, by Entity Diversity Point Statistics

Upon delivery of the decision, protests devolved into riots that resulted in significant destruction of property throughout the study extent. Corresponding protest events are enumerated below, and overlaid onto the tweet collection graph shown in Figure 13.

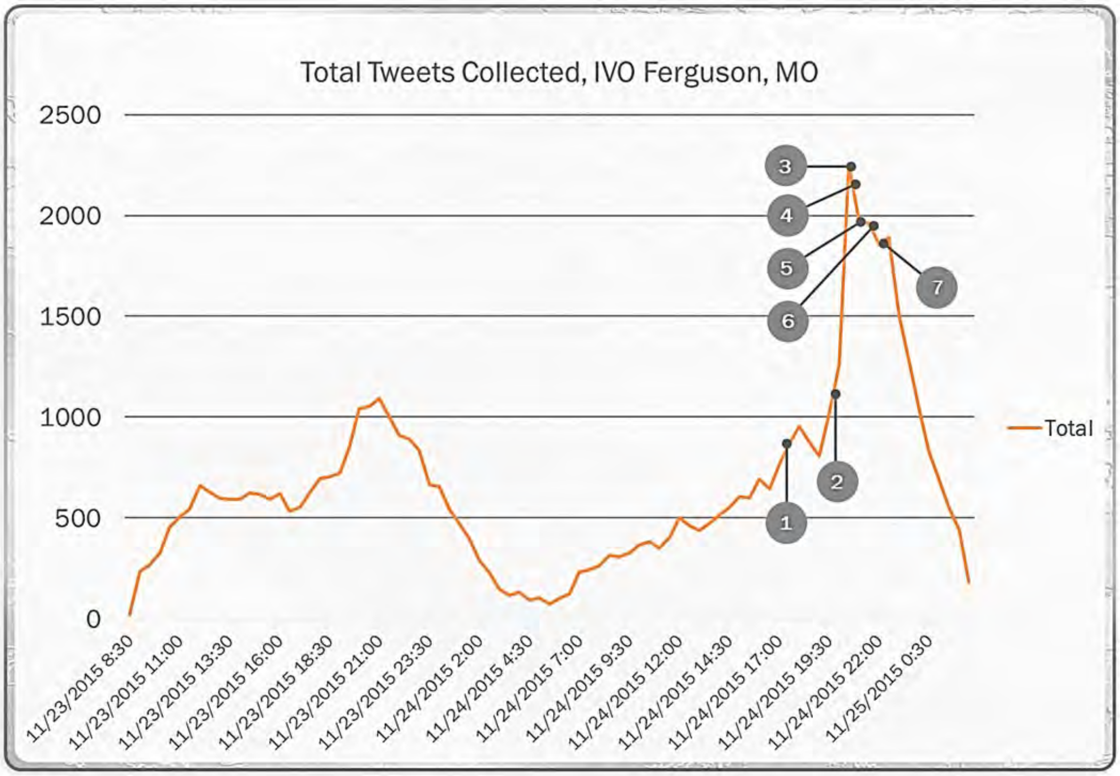

Figure 13 Volume of Tweets per 30 Minute Interval from Ferguson Bounding Box, and Significant Events

1. *17:10,* the office of the prosecuting attorney for Saint Louis, Missouri announces that grand jury's decision will be announced at approximately 20:00.
2. *20:15,* the prosecuting attorney begins his address to the courtroom at the Clayton Justice Center in Saint Louis, Missouri.
3. *20:30,* the prosecuting attorney discusses findings associated with the shooting of Michael Brown, enumerating specific events that preceded Michael Brown's death. Ultimately it is disclosed that the officer responsible Michael Brown's death will not be indicted.
4. *20:40,* the crowd gathered outside the Ferguson Police Department becomes active, and chants such as "No Justice, No Peace," and "We've got to fight" are noted. Around this time objects are thrown in the direction of law enforcement officials standing in front of the Ferguson Police Department.
5. *22:05,* there are multiple reports of shots fired, and initial destruction of property—including an attempt to overturn a police vehicle. The assembled crowd begins to disperse.
6. *21:10,* President Obama issues an address urging calm. Smoke canisters are discharged by law enforcement officials in an effort to disperse the crowd.
7. *21:35 until conclusion of events the following morning,* heavy looting of shops, vehicles and buildings set on fire. No-fly zone goes into effect around Saint Louis.

5.2 Development of a community, or social clusters, using geospatial correlation methods

The initial data pull from the Twitter streaming API occurred between 8:42 am on 23 November and 02:29 on 25 November 2014. This amounted to 55,185 total geotagged tweets from a bounding box that encompassed Ferguson and adjacent townships. Of the 55,185 geotagged tweets collected during this period, 787 unique accounts were noted.

For the sake of comparison to traditional methods of analyzing and spatially displaying Twitter data, all tweets from the original pull were parsed by use of a "#Ferguson" hashtag or any hashtag from the bounding box containing "Ferguson." A word cloud of Ferguson associated hashtags is depicted in Figure 14. The distribution of these #Ferguson geotagged tweets is shown in Figure 15.

Figure 14 Ferguson Associated Twitter Hashtags in a Wordcloud

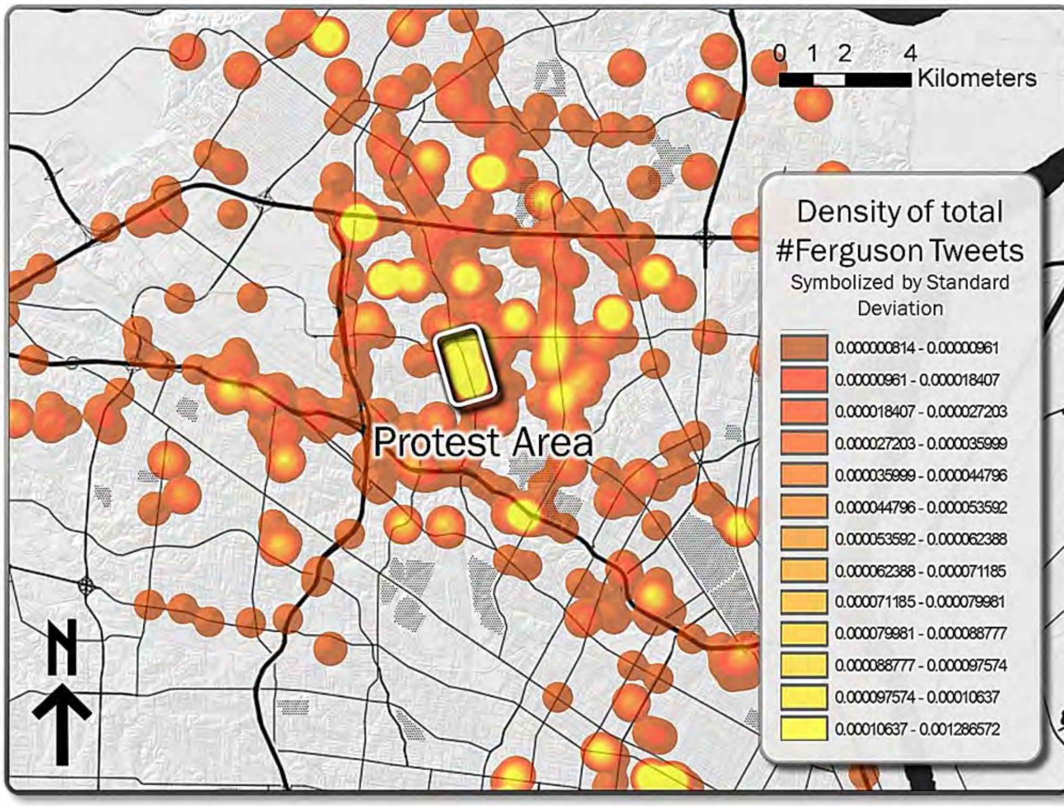

Figure 15 Density of Total Volume of Ferguson Tweets during Protests

All Twitter accounts populating the #Ferguson list were ranked by both total followers and total friends, and each rank was combined to form a numeric designator. This offers a means of masking personally identifiable information, and serves to readily contrast their original rank against ranks derived during subsequent phases of analysis. As an example, "Account_13_4" consists first of the total Twitter followers rank out of 787, and the second is their total friends rank. Aside from a means of comparing methods and coding Twitter user screen names, association with hashtags is not applied elsewhere in the vignette.

The area of interest polygon used to correlate all Twitter accounts from the initial pull to the protest event, which was defined by an area deemed to be the immediate protest area, per documented protest activity on the evening of 24 November 2014. The major attraction for the crowd was the Ferguson Police Department building and Municipal Court.

Using a Python script tool, correlating all account activity to the protest polygon reduced the 787 total accounts to 84. As part of the functionality of this area of interest correlation script tool, a table of these 84 accounts were then joined to the table containing the complete dataset of 55,185 tweets, retaining only those that matched, in order to identify the full range of mobility for Ferguson protest AOI correlated accounts across the complete extent of the collected data. The distribution of tweets from these 84 accounts is shown in Figure 16.

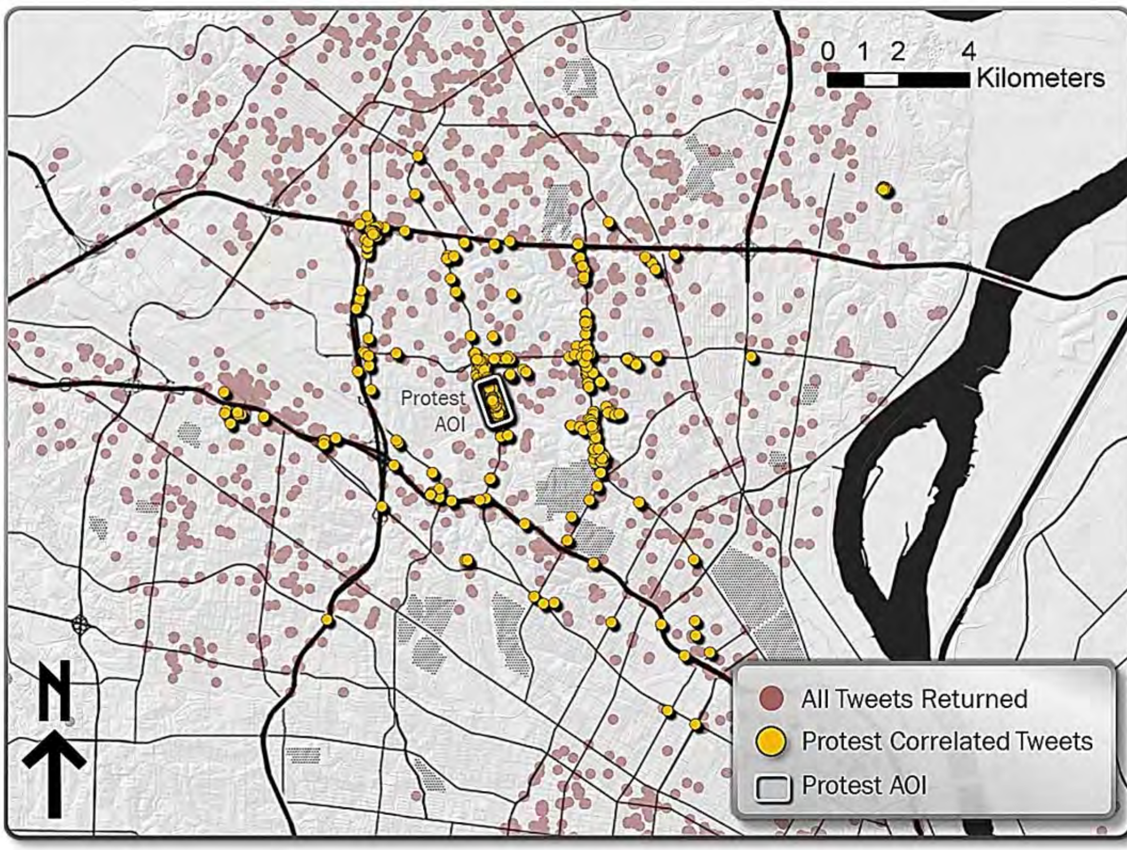

Figure 16 Twitter Events (Tweets) Correlated to the Protest Area of interest (AOI)

5.2. Querying the Twitter Friends and Followers API for Unilateral and Reciprocal Social Media Contacts of Accounts Identified from within Social Clusters

All 84 accounts assessed to be directly participating in the protest, per their correlation to the protest zone polygon, were run through the Twitter Friends and Followers

API as seed accounts, using an API Python script described in a previous chapter. The relationships, or dyads, that resulted from this Twitter Friends and Followers API pull and were used to create an association table with connections that numbered 325,698—specifically 272,708 friend and 52,990 follower relationships. This amounted to 274,100 total unique accounts.

5.3. Isolation of Local Social Network from the Complete Social Network

The area around the protest zone was assessed to develop a locally-appropriate human mobility buffer. As an approximation, the Becker et al. study's distance for median Los Angeles daily weekday range of 5,600 meters was applied as a buffer around the Ferguson Township (Becker et al. 2013). To confirm the fit of this buffer to protest correlated agent behavior, a standard deviational ellipse was calculated from tweets correlated to the protest area of interest polygon. Using this method, the ellipse was found to be predominately inside the Becker derived mobility buffer.

The 325,698 relationships developed from the original 84 protest polygon seed accounts were culled by filtering for only accounts that tweeted from inside the mobility buffer during the night of the protest. The result was a reduction of the 787 total unique accounts populating the hashtag table, to 139 accounts. These 139 accounts consist of both the 84 seed accounts, and 55 additional accounts not correlated to the original protest polygon, but active during the study period, inside the mobility buffer, and either a friend or follower of the original 84 seed accounts. The result is a significantly reduced network that consists only of those relationships assessed to be spatially relevant to Ferguson protest events. The tweets of these 139 accounts are shown in a social network diagram in Figure 17, a depiction of the original AOI correlated tweets and their social connections inside the mobility buffer. Both AOI correlated tweets and social connections, generated during peak protest activity—per Figure 13—are depicted in Figure 18, as a spatiotemporal rendering.

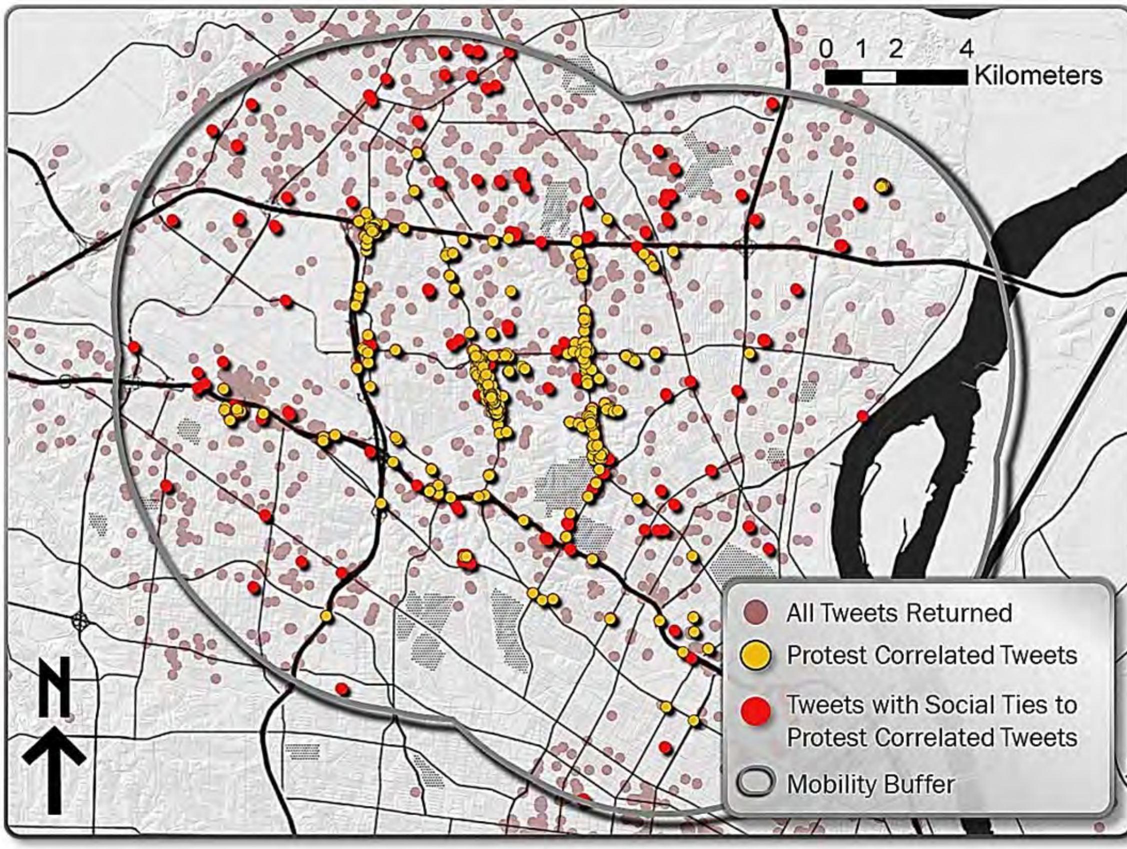

Figure 17 Protest Correlated Tweets and Social Connections

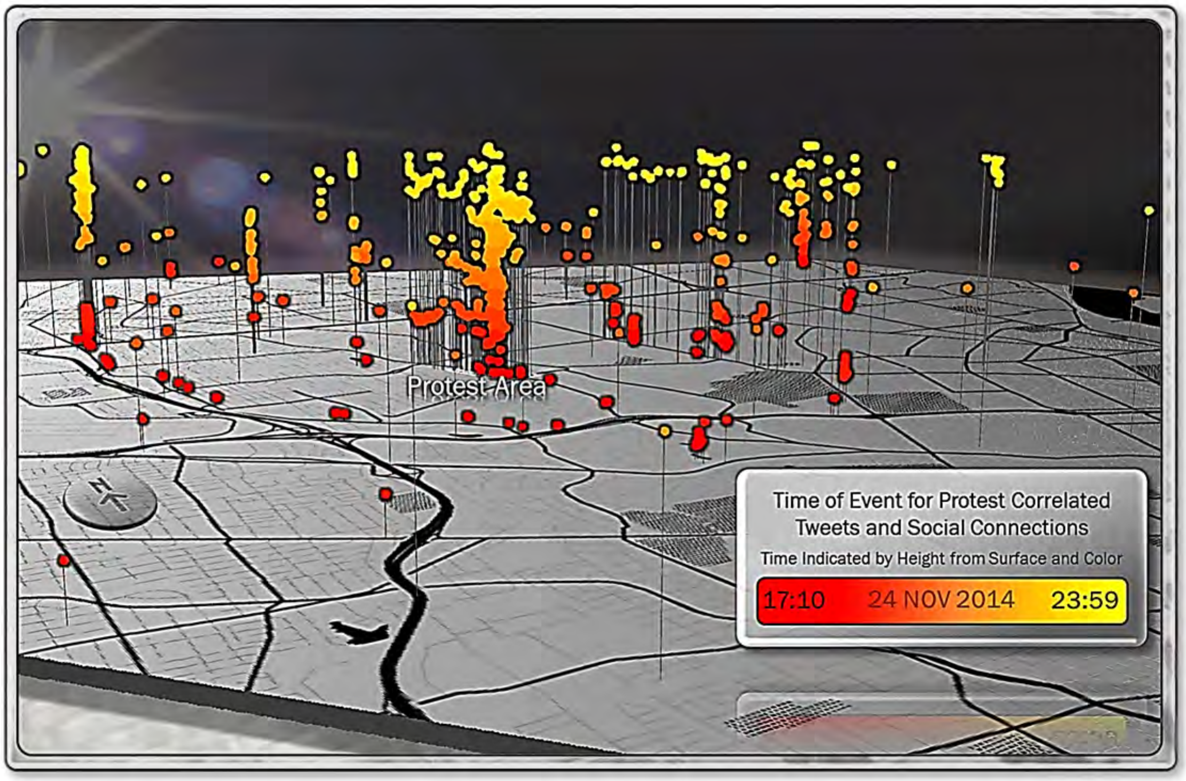

Figure 18 Spatiotemporal Rendering of High Interest Tweets during Peak Protest Activity

5.3.1. Analysis of Twitter Account Activity

Connections between the 139 accounts amounted to 666 relationships on the final spatially constrained association table. This spatially reduced table was ingested into ORA, and run for all social network centrality measures. Figure 19 shows the social network diagram produced by ORA. The resulting table contains all dyads categorized by each of the associated centrality measures for both nodes. As an example of a spatialized social network diagram, many-to-many relationship connections between nodes—as geospatially mobile entities—were rendered as lines, where tweets occurred within the same hour and during peak protest activity. This spatialized social network diagram is depicted in Figure 20.

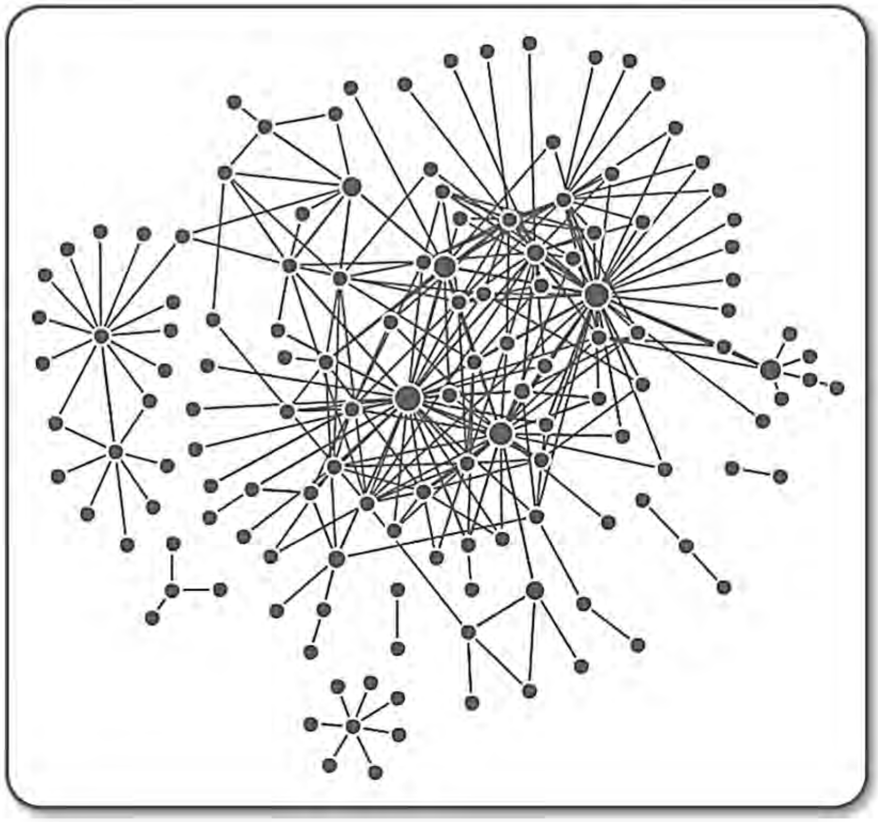

Figure 19 Spatially Reduced Social Network

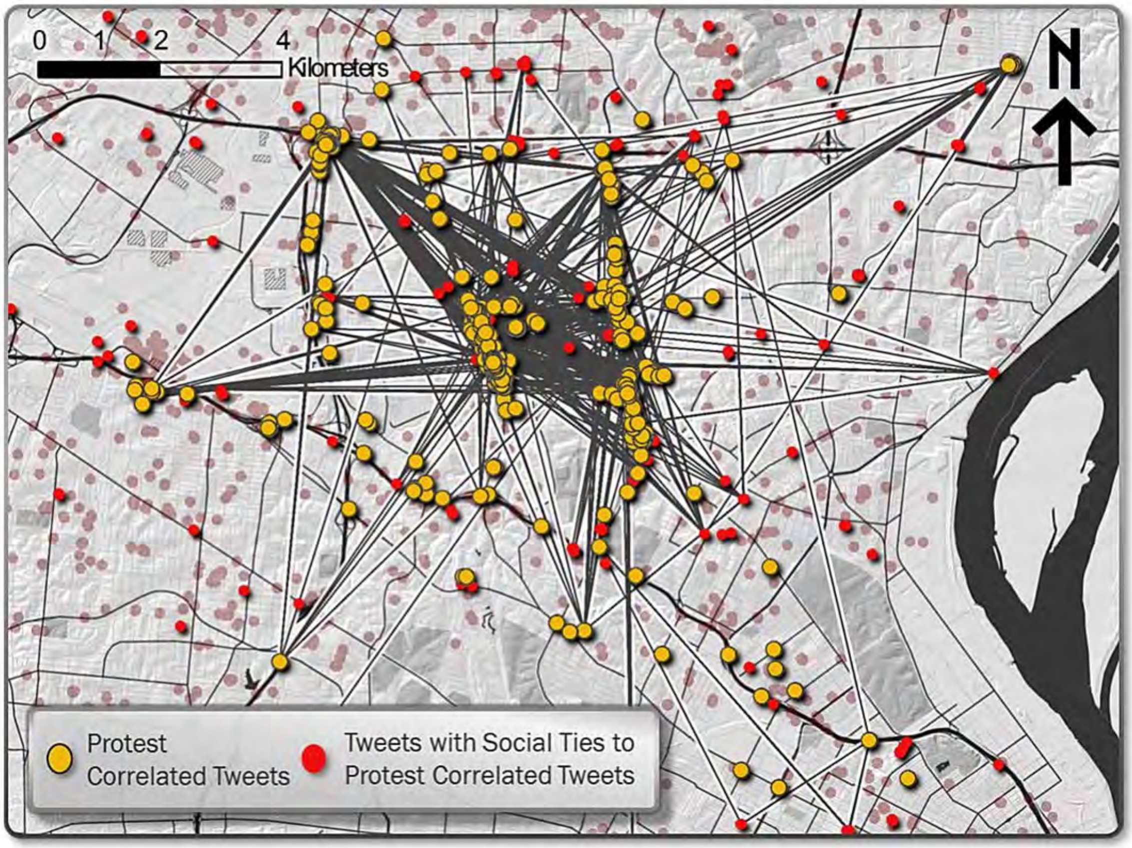

Figure 20 Spatialized Social Network for Same-Hour Coincident Tweets during Peak Protest Activity

Additionally, the Twitter content associated with each of the 139 accounts—photos and text—was reviewed to determine each account's role in protest events. This assessment principally consisted of categorizing each account as either a journalist or not, to determine which centrality measures were most applicable to active protest participants, instead of passive protest participants. Just as the aforementioned Buzzsumo influence assessment process sought to screen journalists by removing accounts with low reply ratios (buzzsumo.com), this process is in part intended to isolate the most locally impactful accounts.

Finally, where available, accounts were assigned a Klout social media influence metric score using an extension that directly appends Klout scores to a Twitter feed. This approach was intended to incorporate a commercial social media influence metric to compliment spatially reduced social network centrality scores.

All of these scores were joined to the Twitter event shapefile as social network metadata—accessible via the shapefile's attribute table. The addition of these values to the shapefile is essential to subsequent steps, including the development of a spatial proximity network using the Blue Starling script tool, and the assessment of socio-spatial correlation. As an example of spatial metadata, social media events were binned by unique user count in Figure 21, and by a combination of this measure, as well as mean betweenness centrality and cumulative time in Figure 22. By using this method, anomalous concentrations of potentially significant activity away from the protest site can be identified.

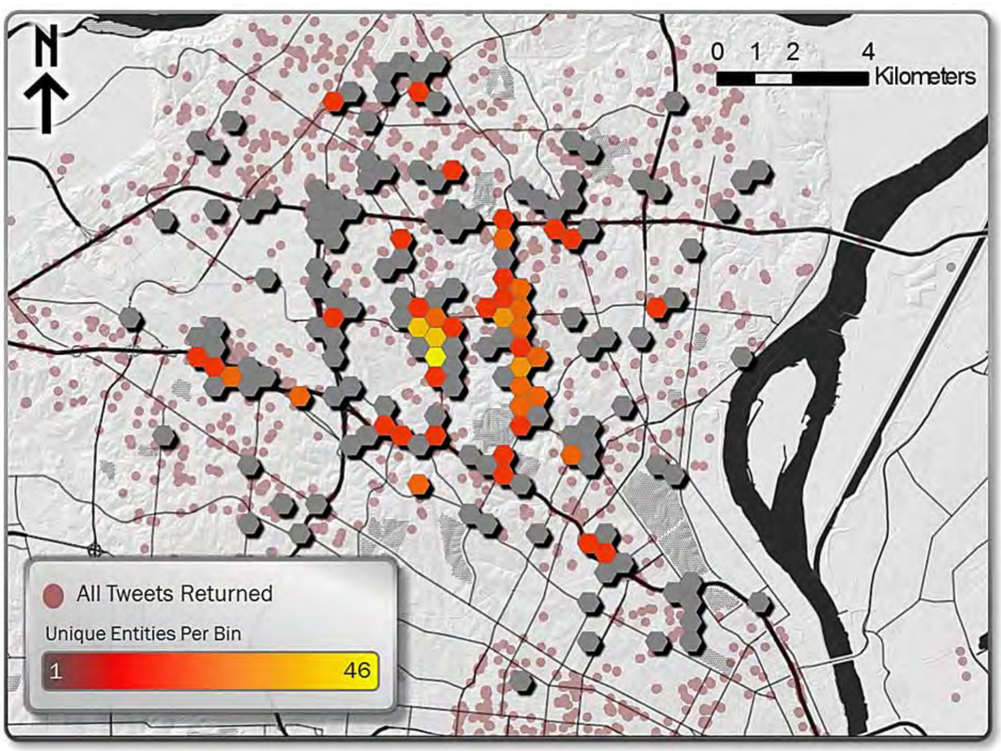

Figure 21 Unique Entities of Interest by Bin

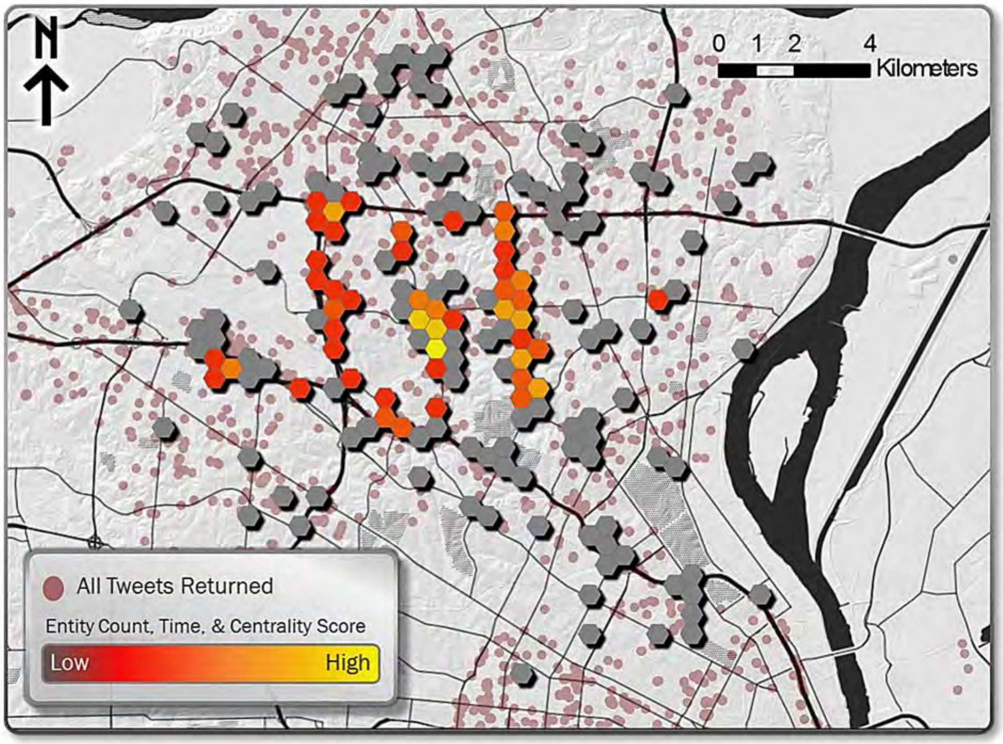

Figure 22 Unique Entities, Mean Betweenness Centrality, and Cumulative Time Score by Bin

5.3.2. Comparative Analysis of Twitter Accounts

The principal intent of this process was to identify discrepancies between traditional influence metrics such as friends, followers and the Klout metric, and the adjusted values associated with entities who have demonstrated through their tweets that they wield local influence—as indicated by an actual role in the protests. Top ranked entities per a multitude of social network scores are included in Table 5, however for the sake of space, all centrality measures developed as a result of the spatial reduction process have been truncated using their average rank.

Table 5 Protest Area Twitter Accounts and Social Connections, by Ranked Social Scores

Account	Journalist	Followers	Friends	Klout	AVG_Rank
Account_21_148	No	7	28	6	3.75
Account_28_21	Yes	10	5	23	4.35
Account_7_271	Yes	2	53	3	6.05
Account_76_540	Yes	24	72	20	11.65
Account_61_337	Yes	20	62	1	12.25
Account_45_257	Yes	16	50	7	12.75
Account_25_26	No	9	6	11	14.2
Account_88_211	Yes	26	41	2	14.45
Account_38_97	Yes	14	13	10	15.45
Account_4_3	No	1	1	8	15.9
Account_34_88	Yes	12	12	15	16.2
Account_44_135	Yes	15	25	35	18.45
Account_242_463	No	51	68	19	21.25
Account_122_82	No	33	11	40	22
Account_14_129	Yes	5	23	16	23
Account_182_287	Yes	42	55	38	23.3
Account_50_229	Yes	17	46	24	23.75
Account_8_193	No	3	35	5	24.2
Account_96_204	Yes	28	40	36	25.3
Account_139_122	No	38	21	21	25.55
Account_59_137	No	19	26	41	25.55
Account_120_319	No	31	60	26	26.1
Account_323_686	No	60	76	30	26.55
Account_726_721	No	78	77	32	28.3

Certain centrality measures favored the popularity of journalists while some favored non-journalists with local connections. As discussed below, after being adjusted to spatial parameters, social network measures most likely to return local activists within the top ranked accounts were clique count, clustering coefficient, cognitive distinctiveness, capability, and Klout score.

Journalists made up the bulk of the top tier of Ferguson hashtag associated accounts, however top accounts from the spatially reduced dataset include more equitable representation from activists and journalists. Account_4_3 was the highest ranked account from the 787 account hashtag list, and this entity was also assessed to be a locally active protest contributor. However, when compared to other active protest contributors from the spatially enhanced dataset, Account_4_3 was not assessed to be as significant. Examining social network measures based on the spatially reduced association table, journalists scored highest in degree centrality, betweenness centrality, and Boniach Power Centrality—a metric that computes centrality based on the centrality of a node's neighbors.

Examination of accounts associated with assessed organizers actively fomenting protests in Ferguson indicated that their highest ranking social network measures did not include any form of degree centrality. Two such notable entities included Account_122_82 and Account_323_686, who did not score well in the overall hashtag rankings and received low Klout scores—ranked 40 and 30 respectively. Account_323_686 which appears to be associated with a benign local organizer scored best in clique count, clustering coefficient, and cognitive distinctiveness. The activity of Acount_323_685 is depicted in Figure 23.

Figure 23 Evidence of Protest Organization Activity after 24 November Ferguson Events by Acount_323_685

Account_122_82, potentially associated with instigating violent behavior during the protest, scored exceptionally well in capability, closeness, and effective network size, including scores among the overall top five. Twitter activity associated with Account_122_82 is depicted in Figure 24.

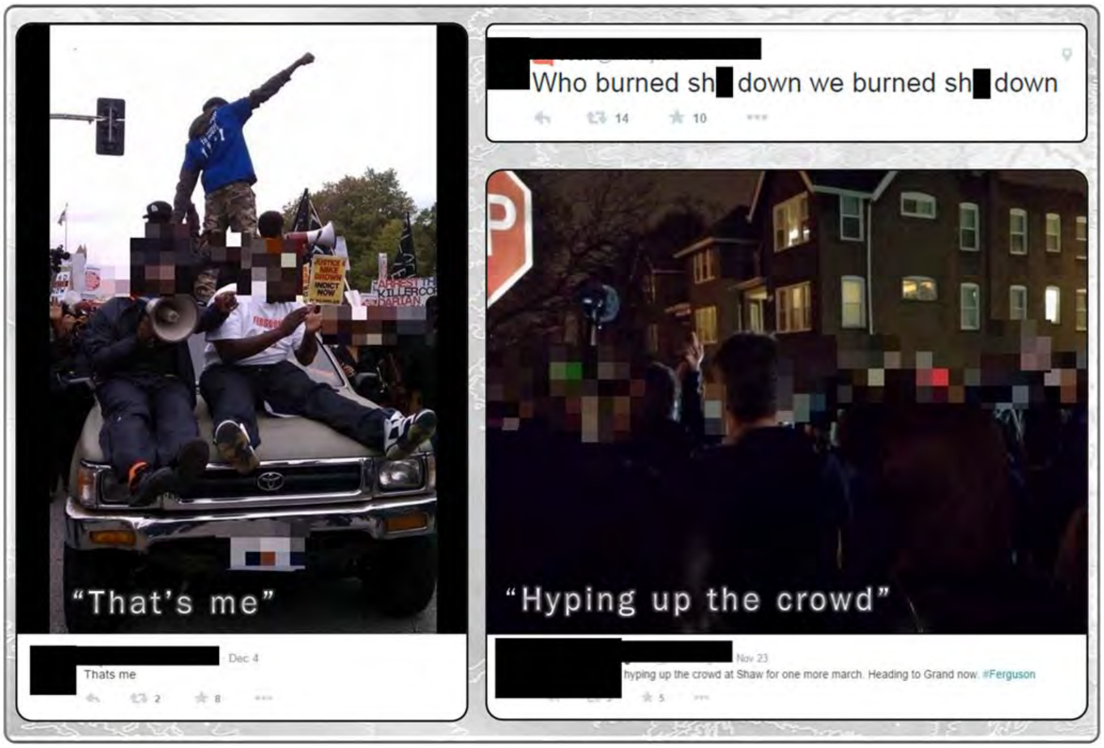

Figure 24 Account_122_82 Seen Actively Contributing to Protests, Indicating "That’s me" As a Caption for the Image on the Left, and "Hyping up the Crowd" for the Image on the Lower Right

Organizers assessed to have a positive effect on protests, those discouraging violence and enabling more peaceful forms of demonstration, were more socially active—from a standpoint of tweet volume. Account_21_148 was not among the top 20 accounts analyzed from the 787 Ferguson hashtag associated accounts list, however when the association table was reduced per spatial criteria, Account_21_148 was a top performer. As depicted in Figure 25, Account_21_148’s actions include running a safe house to shelter protestors when events turned violent, through a charitable organization. Using geotagged tweets, this account was verified at the safe house after having been on site at the protest, as depicted in Figure 26, where tweeted photos of the protest were geospatially corroborated.

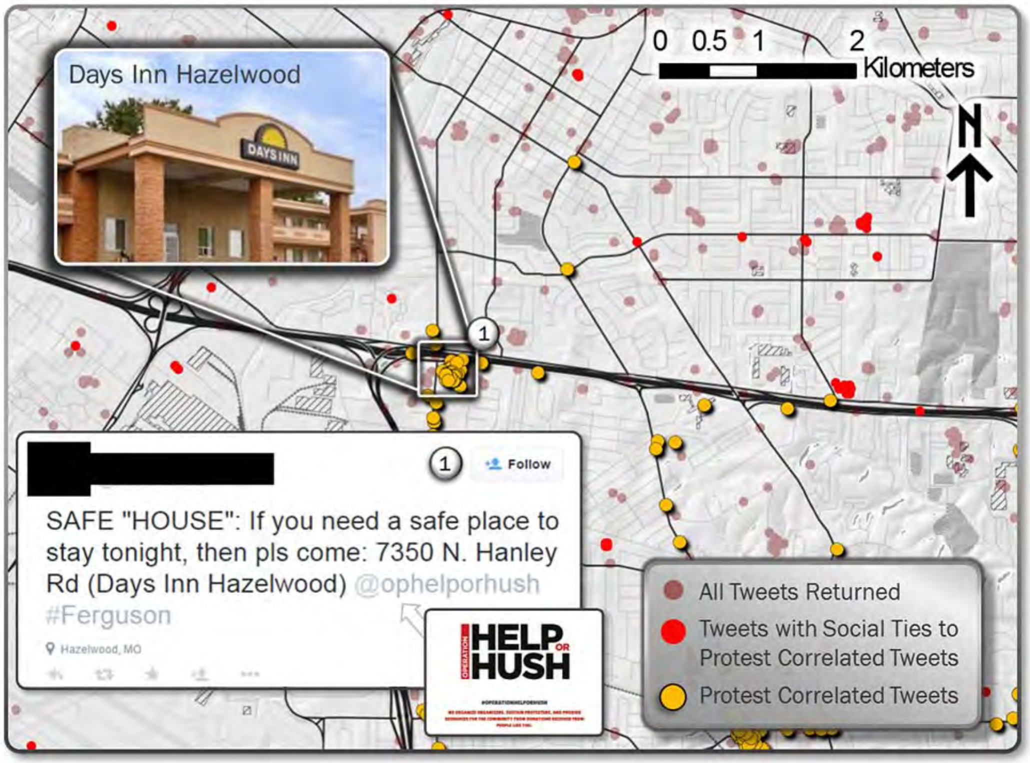

Figure 25 Account_21_148 Associated Protest Safe House, Associated with Charity "Help or Hush"

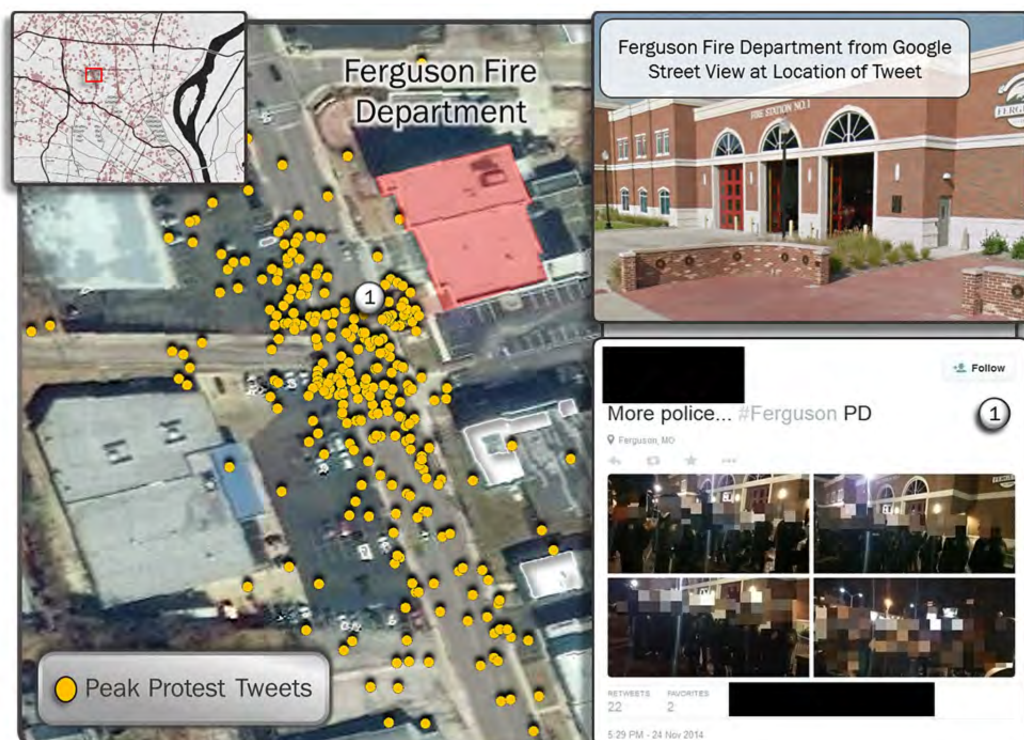

Figure 26 Account_21_148 Verification of Activity at the Protest

The safe house in Figure 25, is also in one of the aforementioned high scoring hex bins depicted in Figure 22. As an aside, the safe house location and other similar dwell locations that could be derived from this process, have the potential to serve as points of interest that factor into an alternative analytic model. It is to be noted that while Account_4_3, the highest scoring account from the 787 account hashtag list, performed better in terms of total degree, Account_21_148 outperformed Account_4_3 in nearly all categories after metrics were adjusted spatially. Account_21_148 also had a superior Klout score from the outset.

5.4. Identification of Socio-Spatial Relationships Using the ESSLVM Blue Starling Script Tool

As mentioned, the centrality measures discussed above were joined to the spatially reduced shapefile which is Twitter event data constrained to the mobility buffer so that social network values could be assessed spatially in subsequent phases of analysis. This shapefile consisted of 139 unique accounts and 1511 social media events. Henceforth this shapefile is referred to as the social network analysis enhanced spatial dataset or, simply, as the SNA enhanced dataset.

Using the previously described Blue Starling Python tool which infers social connections based on spatial, social and temporal variables, the SNA enhanced dataset was assessed for associations indicative of latent or incipient social connections—that would be subsequently added to the association table. Input variables used with the tool included the Twitter user screen name as the unique identifier, the closeness centrality field as the social network measure, and the tweet's date of creation was used to populate the date-time field. Other variables included a radius of 25 meters, an SNA threshold of .001 difference in closeness centrality, a date difference of 30 minutes between events, and a minimum occurrence count of 2—meaning that entities forming a dyad met all other criteria at least 2 times.

5.4.1. Initial Results

Of the 139 unique accounts from the original SNA enhanced dataset, 47 met the specified script tools variable criteria, resulting in 836 total links in the revised association table. When combined with the social network diagram associated with the original association table and the spatially-reduced association table, 85 new links were contributed by the proximity analysis. This network is referred to as the spatial proximity network.

5.4.2. Comparison of Socio-Spatial Relationships to the User Mention-Derived Associations

To compare the results of this process, a separate social network diagram was created from only the user mentions of the 139 total unique accounts assessed. User mentions are considered to be a key indicator of active involvement in a network, and this network interaction has been suggested as a surrogate for network influence (Cha et al. 2010). Additionally, in creating their own latent variable model, the Sadilek study observed that user mentions were likely ideal for the development of unknown social ties, but withheld this data, as user mentions are unique to Twitter datasets, and thus reliance on them in an analytic model would limit that model's application (Sadilek et al. 2012).

Since the SNA enhanced dataset and the resulting proximity dataset are derived from Twitter friends and followers social network content, comparison of the results from the described script tool to the original friends and followers network was determined to be biased. As a result, the social network measures of the spatial proximity network's key entities were compared to those of a network developed separately and exclusively from user mentions—user mentions do not require an API call, as they are part of the Twitter metadata mentioned in Chapter Two. This produced two networks that were disparate, with the former being based principally on spatial associations, and the latter derived from active social network interaction. Additionally, the complete spatial proximity network was sub-divided by occurrence percentile, meaning that different networks were generated from the spatial proximity based on the top 10^{th} and top 20^{th} percentiles, with percentiles determined by frequency of occurrence for each entity pair.

Comparative analysis across multiple key social network entity assessments—using ORA's key entity report function—indicated that top performing accounts were similar between the user mention and proximity networks. These top performers consisted of the aforementioned Account_21_148 and other likely direct associates (as indicated by their connection in the user mention network) whose social network scores were not as competitive when ranked against others in the friends and followers network. These other user mention and spatial proximity network top performing entities included Account_726_721 and Account_515_522, who were also poorly ranked using the hashtag derived list. Additionally, while journalists were often among top scorers in the overall spatial proximity network (likely by virtue of attraction to unfolding events), their significance was reduced when the network's frequency of occurrence was reduced to the top 10^{th} and 20^{th} percentiles.

Figures 27, 28, and 29, show three influence metrics used to demonstrate the significance of the aforementioned entities. Each line is representative of a top entity across all four networks. Metrics shown include spatial proximity (proximity), spatial proximity top 20^{th} percentile (Prox_20), spatial proximity top 10^{th} percentile (Prox_10), and the user mention network (User_Mention). The aforementioned accounts of note

(Account_21_148, Account_726_721, and Account_515_522) are annotated where relevant in each of these charts.

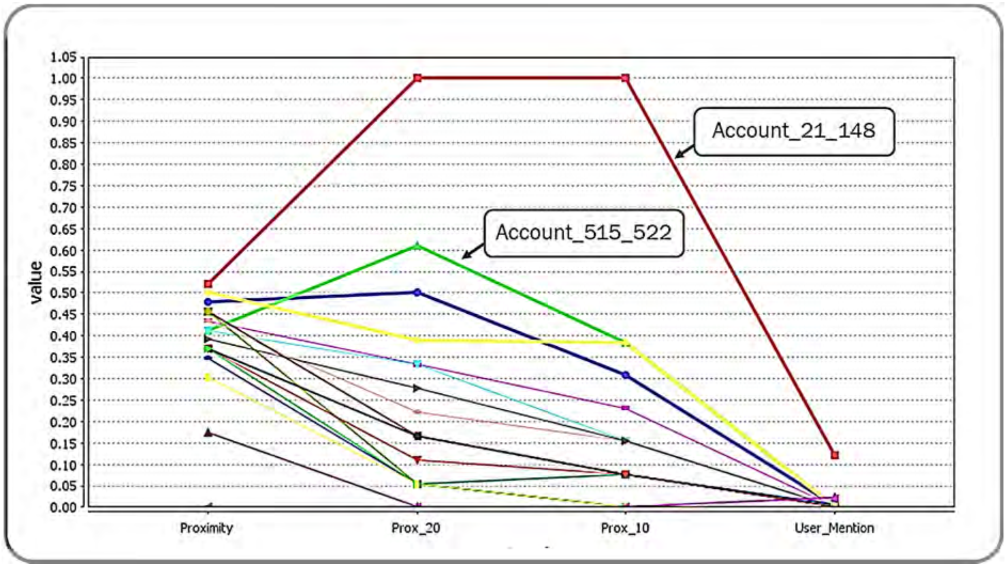

Figure 27 Emergent Leader (Cognitive Demand) Top Twenty Nodes across Four Different Networks

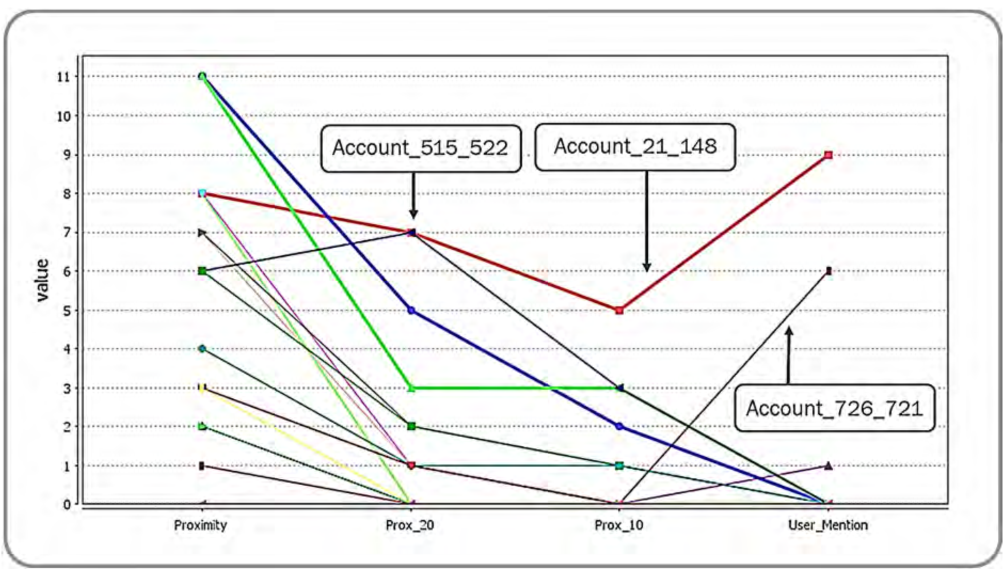

Figure 28 Number of Cliques Top Twenty Nodes across Four Different Networks

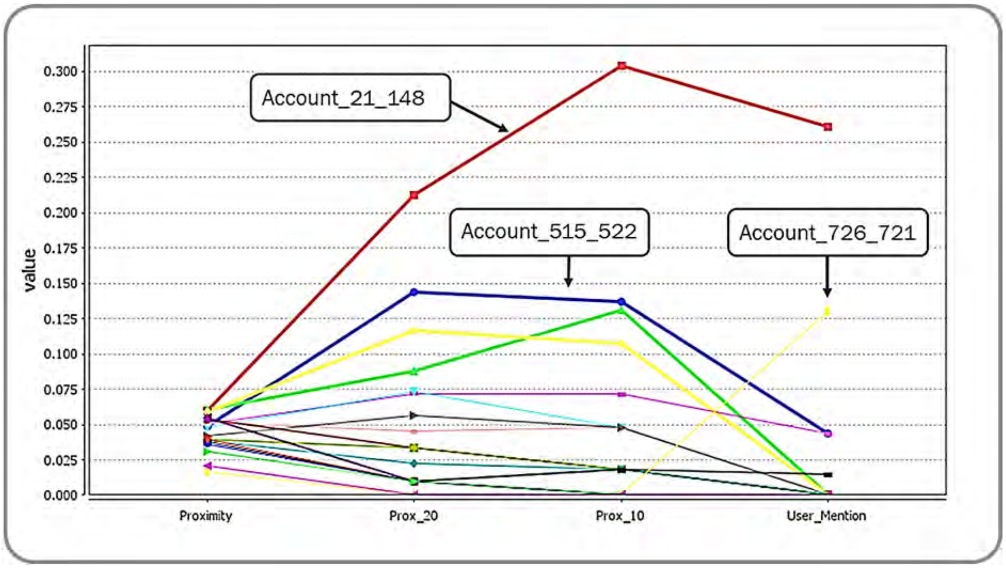

Figure 29 Group Awareness (Shared Situation Awareness) across Four Different Networks

5.5. Additional Measures to Cull the Network

Using the shortest path matrix script, spatially derived relationships were screened to include only those that were within three steps of another node. This screening process resulted in another network with the highest overall network density, effectively reducing the total spatial proximity network by 27%.

5.6. Analysis of the Correlation between Social Network Accounts Centrality Measures and their Geospatial Distribution

A key theoretical component of this thesis is the advancement of propinquity effect as a concept that can be applied to individual agents in support of social network analysis. For this hypothesis to be validated, it was posited that there must exist statistically significant patterns indicative of the correlation of distance between individual entities, and their social network scores.

The Blue Starling spatial proximity association script tool used to create the proximity dataset was rerun, except the distance threshold was set to fifty meters, per observations from the Brieger study on social interaction (Brieger et al. 2003). Of all accounts assessed using this method, 26 had an adequate number of events to allow for calculation of the Spearman's Rank Correlation Coefficient which measures the strength of association, or statistical dependence, between two ranked variables. In this analysis, the rank order of different centrality scores and distance was evaluated for each dyadic set of the 26 participating accounts—those accounts with enough spatial proximity occurrences to qualify for the process. Ranks were assigned in ascending order, so a positively correlated score is indicative of the correlation of lower social network scores (those that are less significant) to shorter (lower) distances, and a negatively correlated score is indicative of correlation of higher social network scores to lower distances. This positive correlation likely means that an entity is surrounded by low centrality or repulsed by high centrality scores, and negative correlation means that an entity surrounds itself or is attracted to higher centrality scores. Using the SciPy Spearman Correlation Coefficient Python script, all accounts were run recursively for each measure, for a total of 442 total correlation runs. Sixteen of those 26 accounts resulted in correlation coefficients with passing p-values (< 0.05). An example of a positively correlated score is provided in journalist Account_11_120, and a negatively correlated score in non-journalist Acccount_25_26 as illustrated in Figures 30 and 31. These figures were created using a Python script and the Matplotlib Python library.

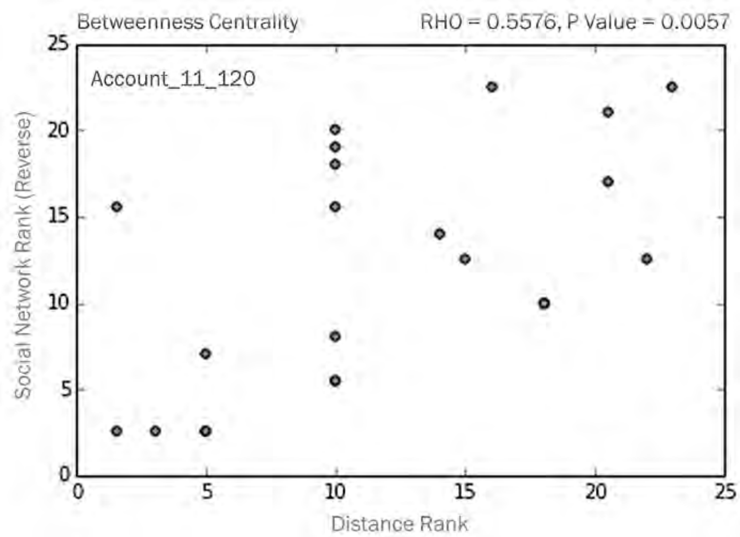

Figure 30 Betweenness Centrality to Distance Correlation for Account_11_120

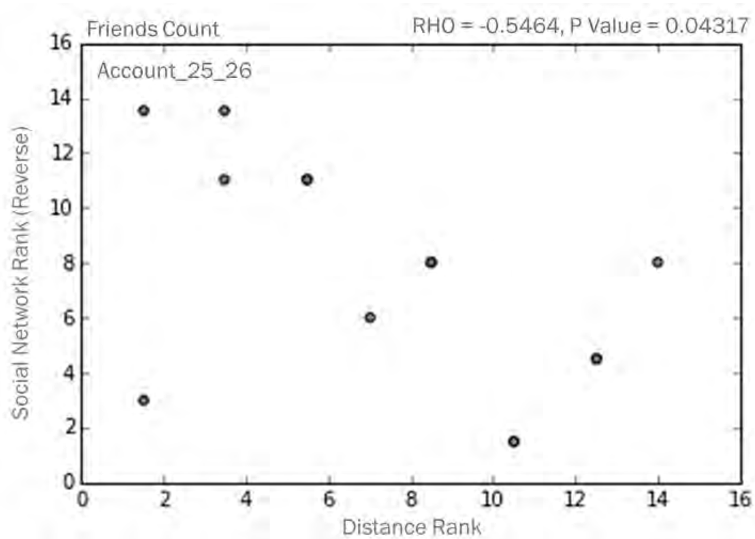

Figure 31 Twitter Friends Count to Distance Correlation for Account 25_26

Non-journalists with passible Spearman coefficients who demonstrated statistically significant spatio-social correlation to other entities, included accounts previously mentioned in Section 3.2.1, such as Account_21_148, Account_515_522, Account_726_721, and Account_25_26. Per analysis of the user mention network, these accounts are likely associates. Additionally, after performing Newman Grouping community detection in ORA (see Figure 32), it was determined that these accounts also comprised a distinct segment of the overall network—further solidifying the connection between social and geospatial realms. While different high-scoring non-journalists had different roles in the event, from a social network standpoint (and in many cases also geospatially) they are interconnected.

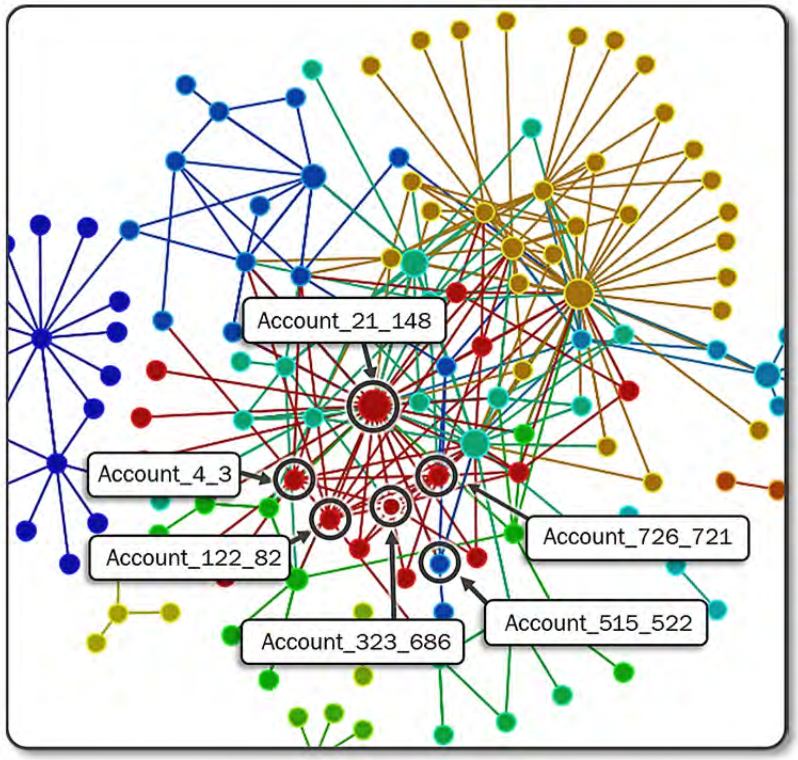

Figure 32 Newman Grouping (Groups Symbolized by Color) of Spatially Reduced Social Network Graph, Illustrating Connections between Nodes of Interest

The four non-journalist accounts that met statistical criteria were both positively and negatively correlated, suggesting that they were either surrounding by other high scoring entities, or low scoring—both of which could have bearing on protest organization. In most cases however, they did not have an especially strong Spearman Correlation coefficient (rho) value, though when symbolized by percentile, these values did exceed median values—where percentile was determined per centrality measure or a social metric for all entities. Notably, for every entity with a passing p-value, correlation was either consistently positively or negatively correlated across a range of 17 social network centrality measures.

Additionally, as depicted in Figure 33, certain journalists also demonstrated strong positive or negative correlation, which from a practical application standpoint, would be ideal for identifying which correspondents have access and can serve as sources most apt to monitor the actions of suspect entities.

In Figure 33, green boxes indicate the strongest positive correlation and red boxes indicate negative correlation. Yellow is indicative of no correlation. Shades between these colors indicate values between the endpoints and the middle. To the left of each value, a green circle indicates that the p-value for that measure was passing (< 0.05) and red indicates a non-passing p-value (>0.05). Centrality measures shown in Figure 33 include the following: (A) Followers Count, (B) Friends Count, (C) Boniach Power Centrality, (D) Capability Centrality, (E) Closeness Centrality, (F) Clustering Coefficient, (G) Cognitive Distinctiveness, (H) Eigenvector Centrality, (I) Out-Degree Centrality, (J) Total Degree Centrality, (K) Betweenness Centrality, (L) Klout Social Media Influence Metric, (M) Effective Network Size, (N) Hub Centrality, (O) In-Degree Centrality, (P) Clique Count, (Q) Authority Centrality. The complete output is available in Appendix A.

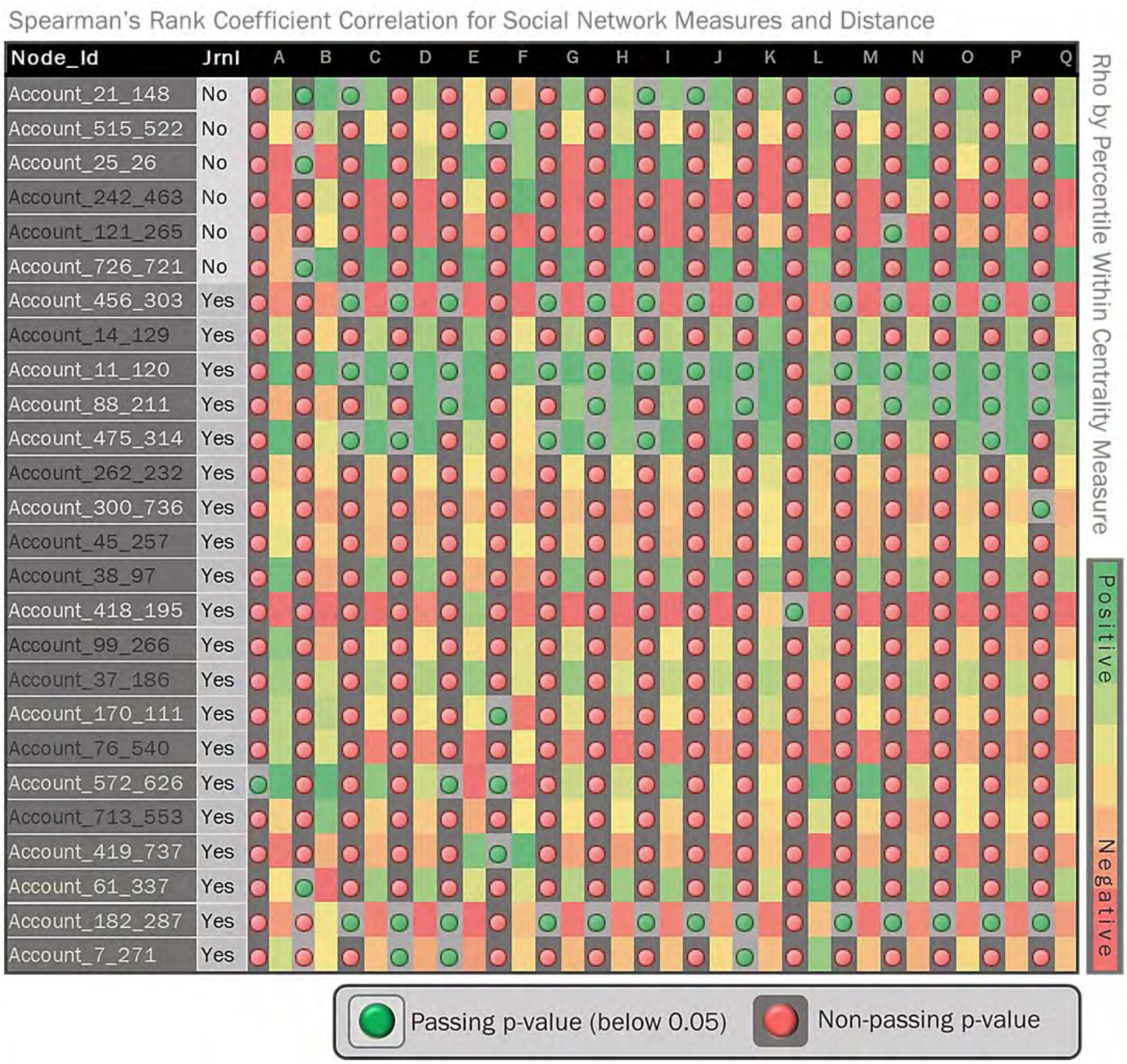

Figure 33 Spearman's Rank Correlation Coefficient for Social Network Measures and Distance, including all Accounts and Highlighted Accounts as those that had Passing p-values. The complete results are available as comma-delimited text in Annex A

CHAPTER 6. REVIEW AND CONCLUSIONS

This study's main impetus was proving the conceptual viability of enhancing social network analysis via geospatial means. A major benchmark for this improvement, was the identification of key influencers that would otherwise be obscured by ambient entities, if traditional non-spatial or existing minimally spatial social network analysis methods were applied. Commensurately, a key supporting task was the inference of social connections using spatial, social, and temporal variables.

At each methodological level this study achieved a degree of success. The analysis of spatially reduced entity social network scores resulted in the isolation of potentially significant influencing actors not detected using metadata or commercial influence metrics; the Blue Starling script tool returned new connections in part corroborated by comparison to the Twitter user mention network, and the correlation of socio-spatial factors had statistically meaningful returns that were also commensurable with other steps comprising the complete ESSLVM system.

It is assessed that this study achieved its core research objectives, and validated the hypothesis that position in a social network has bearing on an individual's relationship with others in physical space, and as a result, individuals or organizations postured to influence a network via direct conduits such as local leadership figures and on-site organizers, possess a qualitative advantage. Additionally, because there exists a reciprocal relationship between an individual's position in a social network and their position among others in physical space, geospatial assessment techniques can be used to infer social connections.

6.1. Vignette Observations

The hashtag parsed Twitter dataset was spatial by most social media analysis measures—social media data was extracted by placing a bounding box around the area of interest. Top performers within this dataset were not representative of those entities assessed to be most actively fomenting protest behavior. By correlating social media accounts to a specific area, and constraining their social media relationships to a surrounding mobility buffer, actors most likely to influence local events emerged.

In the context of protest events, the ESSLVM Python script suite underscored entities of influence and others ostensibly responsive to this influence. Spatial proximity association methods successfully augmented the Twitter friends and followers social network with new edges representative of unknown or incipient connections. This approach was in part validated by comparing top social network scores from the spatially derived proximity network to the Twitter user mention network. The merit of socio-spatial connections was also explored through the correlation of entity social network values to the distance between entities. Beyond buttressing the prediction of social ties using geospatial analysis, socio-spatial correlation results serve to substantiate this study's chief hypothesis and are likely to significant to other methods used to quantify influence and interpersonal dynamics.

Outlined ESSLVM analytic modalities identified many low or mid-level accounts ostensibly exhibiting an ability to wield local influence, which would not have been noted among a list of top scorers derived from many existing social media influence metrics. Use of the Klout social media metric however, did prove valuable where local influencers were exceptionally socially active—as was the case for Account_21_148—though this account's associates were only isolated during subsequent ESSLVM spatial analysis.

Using these results, social unrest could be mitigated through positive engagement with potential non-violent organizers such as Account_21_148 and Account_323_686. Possible riot instigators such as Account_122_82 could also be engaged through means most conducive to the resolution of destructive events, such as those that occurred in Ferguson. Entities who are consistently attracted to influencers, but not influencers themselves—per socio-spatial correlation analysis—could be assessed for human source contact operations, whereby cooperative human sources are used to identify information of value.

6.2. Comparison to Existing Latent Variable Analytic Models

The premier analytic model reviewed was Sadilek's FLAP (Sadilek et al. 2012). ESSLVM does not likely threaten FLAP outright as a systemic means of inferring social connections through geospatial analysis, however ESSLVM's spatial reduction, configurable spatial proximity association, and socio-spatial correlation were not covered in FLAP, and could stand to improve FLAP's latent variable identification. Additionally, ESSLVM's theme of influencer identification was not intrinsic to FLAP, and as such is representative of a novel contribution.

ESSLVM also likely offers some modicum of improvement over the Crandal model (Crandal 2012), which—as emphasized by Sadilek—was likely overly reliant on geospatial collocation as a means of inferring social connections. ESSLVM adds to geospatial collocation other means of assessing contact as indicative of a social relationship, and screening coincidental events.

The spatially reduced social network was also adapted for processing via ORA's "Detect Spatial Patterns" and "Geospatial Assessment." However as a geospatial analysis platform, ORA did not seem well suited to agent analysis at this scale—grouping entities by locations that did not allow for the desired examination of close-quarters, inter-entity dynamics.

6.3. Future Work

Projected ESSLVM tasks focus on refining the model by incorporating alternative analytic methods, and through testing it against different vignettes. The ultimate intent is to create a deployable application that can be directly compared against other latent variable and social network analysis models cited in this study's literature review.

6.3.1. Network Reintegration

Regarding analysis of the results, since only a fraction of all Tweets are geotagged, an additional measure could include layering non-geotagged tweets from the original unadulterated social network atop those accounts deemed most influential per the complete ESSLVM process. This way, un-located sources of influence who have bearing on locally relevant geo-tagged actors can be incorporated into an overall assessment. These entities could include figures not active on-site, who share social relationships with multiple ESSLVM identified influencing entities. As it applies to the vignette, alternatively identified (known) geotagged and non-geotagged protest influencers could be compared to ESSLVM results.

6.3.2. Semantic Analysis

Another measure that could serve to reinforce many of this study's assessments would be the comparison of observed patterns to those apparent via text classification and sentiment analysis. These two measures could be incorporated into the socio-spatial correlation model to determine whether distance from influencing entities alters behavior.

Semantic analysis could ultimately be used in lieu of the manual examination of social media account activity addressed in the vignette.

6.3.3. Alternative Practical Use Cases

The vignette used here was a short duration protest event, which proved practical in terms of the volume of information processed, and the intervals at which this information was generated. From the standpoint of narrative, this vignette was also relevant as a result of the influence themes that were manifest. Future work however, would ideally focus on a larger geospatial extent, include more data, and cover a bigger temporal span—though by virtue of the model's endemic nature, the study extent should not greatly exceed a city-sized environment. This would allow for the analysis of patterns or life, and entails having adequate social media events to identify residences, places of business and routes of travel. An example would be the identification of points of interest associated with a criminal enterprise in a particular city. This type of work would also allow for commensurable comparison to other latent variable models reviewed. Additionally, as covered in the introduction, ESSLVM's commercial applications stand to be explored.

6.3.4. Creation of a Stand-Alone or Minimally Dependent Application

While this study made use of an individually developed Python script tool, the process remains specialized, and requires use of client-side social network analysis software. Future versions of this tool would ideally incorporate NetworkX to a greater extent. NetworkX was used to calculate a social network shortest path matrix, however complete integration will allow a user to execute such a study with fewer dependencies. Using NetworkX, additional functionality allowing for control over social network parameters that define socio-spatial relationships could be directly embedded into the ESSLVM script tool. In this form, ESSLVM would also benefit from an algorithmic process tied to script variables, which also links now disparate stages of analysis, to glean additional statistical insight. This more monolithic application would also ideally feature a GUI that could be accessed outside of ArcGIS Desktop.

6.4. Summary

ESSLVM is intended to reduce network uncertainty and identify key influencers in a manner that improves upon existing analytic processes by geospatially decomposing nebulous social media networks into locally relevant networks, wherein tangible results are more likely. The efficacy of ESSLVM is evidenced in the satisfaction of research objectives, and ultimately, an output in which entities deemed to be influential—per social media content analysis—were ranked higher than when using alternative processes. While many elements of ESSLVM are automated, it is still principally an analytic test-bed that will require development into a deployable application. Initial results however, indicate that ESSLVM shows great promise as a geospatially-enabled, social network analysis application, and that it would benefit from a diverse range of additional test cases.

REFERENCES

Anger, Isabel, and Christian Kittl. 2011. "Measuring Influence on Twitter," Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies, Graz, Austria, September 7-9.

Augure. 2015. Influencers strategy: How to Attract and Engage your Key Influencers." *Augure*. Accessed March 1, 2015, http://www.augure.com/resources/whitepapers/guide-influencers-strategy.

Backstrom, Lars, and Jure Leskovec. 2011. "Supervised random walks: Predicting and recommending links in social networks." Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11), Kowloon, Hong Kong, February 9 - 11.

Backstrom, Lars, Eric Sun, and Cameron Marlow. 2010. "Find Me if You Can: Improving Geographical Prediction with Social and Spatial Proximity." Proceedings of the 19th International Conference on World Wide Web, Raleigh, USA, April 26-30, 2010, 61-70.

Bakshy, Eytan, Jake Hofman, Winter Mason, and Duncan Watts. 2011. Everyone's an Influencer: Quantifying Influence on Twitter. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11), Kowloon, Hong Kong, February 9 – 11, 65-74

Becker, Richard, Ramón Cáceres, Karrie Hanson, Sibren Isaacman, Ji Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. 2013. Human Mobility Characterization from Cellular Network Data. *Communications of the ACM* 56, (1): 74-82.

Berger, J. M., and Jonathon Morgan. 2015. *The ISIS Twitter Consensus, Defining and Describing the Population of ISIS supporters on Twitter*. Washington, D.C.: The Brookings Institution.

Bohdanova, Tetyana. 2014. "Erratum to: Unexpected revolution: The Role of Social Media in Ukraine's Euromaidan Uprising." *European View* 13, (2): 347-347.

Bonito, Joseph A., Norah E. Dunbar, Karadeen Kam, Jenna Fischer, Judee K. Burgoon, and Artemio Ramirez Jr. 2002. "Testing the Interactivity Principle: Effects of Mediation, Propinquity, and Verbal and Nonverbal Modalities in Interpersonal Interaction." *Journal of Communication* 52 (3): 657-677.

Burris, Brandon. 2013. *Methods and systems of aggregating information of social networks based on changing geographical locations of a computing device via a network.* Texas/USA Patent US 2013/0268558 A1, filed 2013, and issued 10/Oct/2013.

Can, Fazli, editor, Tansel Özyer editor, and Faruk Polat editor. 2014. *State of the Art Applications of Social Network Analysis*. Cham: Springer.

Carly, Kathleen, Jurgen Pfeffer, Jeff Reminga, Jon Storrick, and Dave Columbus. 2013. *ORA User's guide 2013*. CMU - ISR - 13 - 108 ed. Pittsburgh, PA 15213: Carnegie Mellon University; 1-1288.

Cha, Meeyoung, Hamed Haddadi, Fabricio Benevenuto, and Krishna Gummadi. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. Proceedings of 4th International AAAI Conference on Weblogs and Social Media, George Washington University, Washington, DC, May 23-26, 1-8.

Cheng, Zhiyuan, James Caverlee, and Kyumin Lee. 2010. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. Proceedings of the 19th ACM International Conference on Information and Knowledge Management, Toronto, ON, Canada, October 26 – 30, 759-768.

Cho, YS, A. Galstyan, PJ Brantingham, and G. Tita. 2014; 2013. "Latent Self-Exciting Point Process Model for Spatial-Temporal Networks." *Discrete and Continuous Dynamical Systems-Series B* 19, (5): 1335-1354.

Claypoole, Theodore F. 2014. Privacy and Social Media. *Business Law Today*: n.p.

Cole, Juan. 2014. "Top Attempts by Dictators to Shut Down Twitter in Mideast (Including Turkey's PM Erdogan)." *Juancole.com.* Accessed February 1, 2015. http://www.juancole.com/2014/03/dictator-attempts-including.html.

Crandall, David J., Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, Jon Kleinberg, and Ronald L. Graham. 2010. "Inferring Social Ties from Geographic Coincidences." *Proceedings of the National Academy of Sciences of the United States of America* 107 (52): 22436- 22441.

Cuthbertson, Anthony. 2015. Anonymous Lists 9,200 Twitter Accounts Linked to Islamic State after Hacktivist Collaboration. *International Business Times*, 16 March 2014, 2015, sec 2015.

Davis, Darcy, Ryan Lichtenwalter, and Nitesh V. Chawla. 2013. "Supervised Methods for Multi-Relational Link Prediction." *Social Network Analysis and Mining* 3 (2): 127-41.

DARPA. 2015. "Social Media in Strategic Communication (SMISC)." *DARPA.* Accessed April 2. http://www.darpa.mil/opencatalog/SMISC.html

Division of Behavioral and Social Sciences and Education, Committee on Human Factors, Board on Human-Systems Integration, and National Research Council. 2003. "Dynamic Network Analysis." In *Predictability of Large-Scale Spatially Embedded Networks" Dynamic Network Analysis*, 313-323. Washington, DC: National Research Council

Elliot, J. 2014. "5 Outstanding Social Media Campaigns." *Hallaminternet.com.* Accessed April 25, 2015, https://www.hallaminternet.com/2014/5-social-media-campaigns/.

Esri. 2014. “Tapestry Segmentation.” *Esri.* Accessed May 20. http://www.esri.com/landing-pages/tapestry

Facebook. 2015. “The graph API.” *Facebook.* Accessed April 25. https://developers.facebook.com/docs/graph-api

Freeman, Linton C. 1979. Centrality in Social Networks I. Conceptual Clarification. *Social Networks* 1 (3): 215.

Furht, Borivoje. 2010. *Handbook of Social Network Technologies and Applications*. New York: Springer.

Gauthier, TD. 2001. Detecting Trends Using Spearman's Rank Correlation Coefficient. *Environmental Forensics* 2 (4): 359-62.

Gertz, Bill. 2015. “Special Ops Targets Social Media.” *The Washington Times*. Accessed March 22, 2015. doi. March 18, 201.5

Giordano, Alberto, and Tim Cole. 2011. “On Place and Space: Calculating Social and Spatial Networks in the Budapest Ghetto.” *Transactions in GIS* 15 (s1): 143-70.

Goebbels, Joseph. 1934. “Erkenntnis und Propaganda.” In *Signale der neuen Zeit. 25 ausgewählte Reden von Dr. Joseph Goebbels*, 3. Munich; Zentralverlag der NSDAP.

Guo, Diansheng. 2009. “Flow Mapping and Multivariate Visualization of Large Spatial Interaction Data.” *IEEE Transactions on Visualization and Computer Graphics* 15 (6): 1041-8.

Hodas, Nathan O., and Kristina Lerman. 2014. “The Simple Rules of Social Contagion.” *Scientific Reports* 4, 4343.

Hoi, Steven C. H. 2011. *Social Media Modeling and Computing*. London; New York: Springer

Hootsuite. 2014. “Case study: The Rockefeller Foundation.” *Hootsuite*. Accessed March 1, 2015. https://hootsuite.com/resources/case-study/reach-is-not-the-same-as-influence-the-rockefeller-foundation-case-study.

Huebner, Bryce. 2014. *Macrocognition: A Theory of Distributed Minds and Collective Intentionality*. Oxford: Oxford University Press.

Kadushin, Charles. 2012. *Understanding social networks: Concepts, theories, and findings*. New York: Oxford University Press.

Khonsari, Kaveh Ketabchi, Zahra Amin Nayeri, Ali Fathalian, and Leila Fathalian. 2010. “Social Network Analysis of Iran's Green Movement Opposition Groups Using Twitter.”

The 2010 International Conference on Advances in Social Networks Analysis and Mining. Odense, Denmark, August 9-11, 414-415.

Klee, Miles. 2014. "Facebook's 'People You May Know': An Interview with Gabriel Roth." *The Daily Dot.* Accessed April 3, 2015, doi: June 25, 2014.

Li, Fenhua Li, Jing He, Guangyan Huang, Yanchun Zhang, and Yong Shi. 2014. A Clustering-Based Link Prediction Method in Social Networks. *Procedia Computer Science* 29 (14): 432-42.

Li, Liangda, and Hongyuan Zha. 2013. Dyadic Event Attribution in Social Networks with Mixtures of Hawkes Processes. Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management, San Francisco, California, October 27 – November 1: 1667-1672.

Lia, Linna, Michael F. Goodchild, and Bo Xu. 2013. "Spatial, Temporal, and Socioeconomic Patterns in the Use of Twitter and Flickr." *Cartography and Geographic Information Science* 40 (2): 61.

Lucente, Seth, and Wilson, Greg. 2013. "Crossing the Red Line: Social Media and Social Network Analysis for Uncoventional Campaign Planning." *Special Warfare* 26, (3): 22-28.

Macmanus, Richard. 2012. "What Facebook May do with Glancee, its Latest Mobile Acquisition." *Readwrite:* Accessed April 15, http://readwrite.com/2012/05/06/what-facebook-may-do-with-glancee-its-latest-mobile-acquisition.

Maher, Shiraz. 2014. "Analyzing the ISIS Twitter storm." *War on the Rocks.* Accessed March 2, 2015, http://warontherocks.com/2014/06/analyzing-the-isis-twitter-storm/.

McCulloh, Ian, Helen (Helen Leslie) Armstrong, and Anthony N. Johnson. 2013. *Social Network Analysis with Applications*. Hoboken, N.J: Wiley.

McHugh, Molly. 2014. "Klout Reveals a New Scoring Algorithm," *Digital Trends*. Accessed April 28, 2015, http://www.digitaltrends.com/social-media/klout-reveals-new-scoring-algorithm-and-the-critics-are-quiet/.

Mennis, Jeremy, and Michael Mason. 2011; 2010. "People, Places, and Adolescent Substance Use: Integrating Activity, Space and Social Network Data for Analyzing Health Behavior." *Annals of the Association of American Geographers* 101 (2): 272-1.

Miller, Harvey J. 2005. "Necessary space - time conditions for human interaction." *Environment and Planning B: Planning and Design* 32 (3): 381-401.

Mok, Diana, Barry Wellman, and Juan Carrasco. 2010. "Does distance matter in the age of the internet?" *Urban Studies* 47 (13): 2747-83.

Notess, Greg R. 2013. "Klout." *Online Searcher* 37 (4): 9.

Ota, Yusuke, Kazutaka Maruyama, and Minoru Terada. 2012. Discovery of interesting users in twitter by overlapping propagation paths of retweets. Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology 3: 274-279

Palantir. 2013. "Playing with Palantir." *Inside the Pentagon's Inside the Army* 25, (27).

Pew Internet & American Life Project. 2014. *Mapping Twitter Topic Networks: From Polarized Crowds to Community Clusters; 2014 SRI R8588-15.*

Pozdnoukhov, Alexei, and Christian Kaiser. 2011. Space-Time Dynamics of Topics in Streaming Text. Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, Chicago, Illinois, November 1 - 4: 1-8.

Rhodes, Christopher J. 2011. "The Use of Open Source Intelligence in the Construction of Covert Social Networks." In *Counterterrorism and Open Source Intelligence* Vol. 2, 159-170. Vienna: Springer Vienna.

Sadilek, Adam, Henry Kautz, and Jeffrey Bigham. 2012. Finding Your Friends and Following Them to Where You Are. Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, Seattle, Washington, February 8 - 12: 723-732.

Sadinle, Mauricio. 2012. *The Strength of Arcs and Edges in Interaction Networks: Elements of a Model-Based Approach.* Pittsburgh, Pennsylvania: Carnegie Mellon University.

Sagl, Guenther, Eric Delmelle, and Elizabeth Delmelle. 2014. "Mapping Collective Human Activity in an Urban Environment Based on Mobile Phone Data." *Cartography and Geographic Information Science* 41 (3): 272-85.

Saito, Kodai, and Naoki Masuda. 2013. "Two Types of Twitter Users with Equally Many Followers." Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagrara, Ontario, August 25 - 29: 1425-1426

Schreck, Tobias, and Daniel Keim. 2013. "Visual Analysis of Social Media Data." *Computer* 46 (5): 68-75.

Shaner, Jeff. 2013. "Smartphones, Tablets and GPS Accuracy". *ArcGIS Resources.* 15, July: n.p.

Skarda, Erin. 2014. "What You Need to Know about the 5 Most Successful Social Media Campaigns for Social Change." *Nationswell. Accessed March 15, 2015.* http://nationswell.com/social-media-campaigns-successful-at-change.

Sreberny, Annabelle, and Ali Mohammadi. 1994. *Small Media, Big Revolution: Communication, Culture, and the Iranian Revolution*. Minneapolis: University of Minnesota Press.

Statista. 2014. “Social Media Marketing Spending in the United States from 2014 to 2019 (in Billion U.S. Dollars).” *Statista*. Accessed April 1, 2015. http://www.statista.com/statistics/276890/social-media-marketing-expenditure-in-the-united-states/.

Sullivan, Bob. 2014. "Police Sold on SnapTrends, Software that Claims to Stop Crime before it starts…by Reading everyone’s Tweets, FB Posts, etc." *bobaullivan.net*. Accessed May 1, 2015. https://bobsullivan.net/cybercrime/privacy/police-sold-on-snaptrends-software-that-claims-to-stop-crime-before-it-starts-by-reading-everyones-tweets-fb-posts-etc.

Takhteyev, Yuri, Anatoliy Gruzd, and Barry Wellman. 2012. “Geography of Twitter Networks.” *Social Networks* 34 (1): 73-81.

Taylor, Zac. 2012. “Panelists Discuss 'Arab Spring' Twitter Revolution.” *McClatchy - Tribune Business News* 2012.

Thomas, Owen. 2014. “How the Internet Will Tell You What to What, Where to Go, and Even who to Date.” *Readwrite.* Accessed January 19, 2015. http://readwrite.com/2013/04/10/anticipatory-systems-artificial-intelligence

Tucker, Patrick. 2014. “The US Military is Already Using Facebook to Track your Mood.” *Defense One*. Accessed March 21, 2015. http://qz.com/229800/the-us-military-is-already-using-facebook-to-track-your-mood/

Twitter. 2015. “Success stories.” *Twitter*: Available from https://biz.twitter.com/en-gb/success-stories.

Wang, Liangshun, and Huajing Fang. 2013. “Stability Analysis of Local Swarms with Attraction / Repulsion Force.” *Journal of Control Theory and Applications* 11, (3): 489-494.

Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. Vol. 8. Cambridge; New York: Cambridge University Press.

Watts, Duncan J., 1971. 2003. *Six Degrees: The Science of a Connected Age*. New York: W.W. Norton.

Weidemann, Chris Donald. 2014. Geosocial Footprint (2013): Social Media Location Privacy Web Map. Ph.D. diss., ProQuest, UMI Dissertations Publishing

Wiil, Uffe Kock. 2013. “Issues for the Next Generation of Criminal Network Investigation Tools.” 2013 European Intelligence and Security Informatics Conference, Uppsala, Sweden, August 12 - 14: 7-14

Wire. 2013. “Tellagence Releases Community to Move Influence Marketing Beyond Follower Counts, Twitter Lists, and Social Scoring.” *Internet Wire*. Accessed March 5, 2015. http://www.marketwired.com/press-release/tellagence-releases-community-move-influence-marketing-beyond-follower-counts-twitter-1840690.htm.

Wong, Kyle. 2014. The Explosive Growth of Influencer Marketing and What it Means for You. *Forbes.* Accessed February 20, 2015. doi; October 9, 2014.

Yamaguchi, Yuto, Toshiyuki Amagasa, and Hiroyuki Kitagawa. 2013. “Landmark-Based User Location Inference in Social Media.” Proceedings of the First ACM Conference on Online Social Networks, Boston, Massachusetts, October 7 - 8: 223-234

Yang, Jaewon, and Jure Leskovec. 2015. “Defining and Evaluating Network Communities Based on Ground-Truth. *Knowledge and Information Systems* 42 (1): 181-213

Yin, Ling, and Shih-Lung Shaw. 2015. “Exploring Space–Time Paths in Physical and Social Closeness Spaces: A Space–Time GIS Approach.” *International Journal of Geographical Information Science* (01/09; 2015/03): 1-20.

APPENDIX A: Socio-Spatial Correlation Full Output

Node_Id,Journalist,Follow_PVAL,Follow_RHO,Friend_PVAL,Friend_RHO,Boniach PVAL,Boniach_RHO,Capability_PVAL,Capability_RHO,Closeness_PVAL,Closeness_RHO,Clustering_PVAL,Clustering_RHO,Cognitive_PVAL,Cognitive_RHO,Eigenvector_PVAL,Eigenvector_RHO,OutDegree_PVAL,OutDegree_RHO,TotalDegree_PVAL,TotalDegree_RHO,Between_PVAL,Between_RHO,Klout_PVAL,Klout_RHO,EN_PVAL,EN_RHO,Hub_PVAL,Hub_RHO,In_Degree_PVAL,In_Degree_RHO,Clique_PVAL,Clique_RHO,Authority_PVAL,Authority_RHO

akacharleswade,No,0.517449556,0.040961453,0.003111988,0.185541748,0.023953039,0.142216514,0.31650111,0.063351081,0.694671108,-0.024846396,0.307495197,-0.064538773,0.081420172,0.109981046,0.295511655,0.066155229,0.023953039,0.142216514,0.049839584,0.123692784,0.257018,0.071666191,0.178814071,0.084959406,0.015952503,0.151690365,0.562290782,0.036671471,0.109755106,0.100990464,0.636383953,0.02992322,0.21494782,0.078386815

AnarchistAnnie,No,0.473456276,-0.047191132,0.333326517,0.063656836,0.966882873,0.002734643,0.742061594,-0.021675703,0.802981547,-0.016431921,0.001709467,0.204407786,0.537422469,-0.040604075,0.655652499,0.029365047,0.966882873,0.002734643,0.847399509,0.01267466,0.184978482,-0.087145081,0.221446282,0.080403845,0.962367267,0.003107777,0.297019865,0.068607523,0.788999368,0.017625156,0.819301136,0.015045931,0.415506878,0.053592583

ChozynBoy,No,0.074562201,-0.491072652,0.043175467,-0.546481131,0.510558194,0.192111123,0.922863549,0.028533531,0.457721673,0

.216265118,0.501979224,0.195950617,0.357012257,-
0.266521535,0.271383207,0.315794424,0.510558194,0.192111123,0.820637325,-
0.066751138,0.108225364,-
0.447935353,0.825406595,0.064950626,0.686087463,0.118702869,0.271383207,0
.315794424,0.808544558,-
0.071326346,0.575198216,0.164050595,0.182040122,0.378505373
dellmaj7th,No,0.440826384,-0.319255903,0.974228681,0.013746435,0.2589226,-
0.453632354,0.2589226,-
0.453632354,0.974228681,0.013746435,0.170792318,0.536110964,0.2589226,-
0.453632354,0.170792318,-0.536110964,0.2589226,-0.453632354,0.170792318,-
0.536110964,0.170792318,-0.536110964,0.974228681,-0.013746435,0.2589226,-
0.453632354,0.974228681,0.013746435,0.170792318,-0.536110964,0.170792318,-
0.536110964,0.170792318,-0.536110964
FCarbonne,No,0.535051133,-0.136341401,0.86801295,-0.036685828,0.116314259,-
0.336589886,0.182672611,-0.287990547,0.278240262,-0.236037384,0.244473277,-
0.252816408,0.221433414,-0.265148253,0.194400354,-0.280758891,0.116314259,-
0.336589886,0.380398126,-0.191894071,0.507073062,-0.145713576,0.182267757,-
0.288245795,0.139302648,-0.317932108,0.050250444,-0.412837214,0.440868019,-
0.168970037,0.340746725,-0.208067213,0.18385462,-0.287247541
HyperboleJ,No,0.673512925,-
0.12940149,0.016022766,0.650690655,0.190019078,0.388131268,0.190019078,0.
388131268,0.090984659,0.487599816,0.235518904,0.353884391,0.190019078,0.3
88131268,0.287057814,0.319637515,0.190019078,0.388131268,0.296232159,0.31
3929702,0.357657986,0.278051128,0.738382189,0.10274063,0.190019078,0.3881

31268,0.190019078,0.388131268,0.430189872,0.239728136,0.287057814,0.31963
7515,0.357657986,0.278051128
AdamRndll,Yes,0.226232322,-0.198294811,0.312105966,-
0.166136569,0.030859595,-0.346191628,0.007886574,-0.419281026,0.038397377,-
0.332868081,0.094839939,-0.271262403,0.005048576,-0.440135224,0.024484264,-
0.359749374,0.030859595,-0.346191628,0.016409311,-0.38198677,0.011684261,-
0.399781385,0.222888521,-0.19970158,0.021466798,-0.367219635,0.009635812,-
0.409482895,0.016409311,-0.38198677,0.02135312,-0.367517786,0.024484264,-
0.359749374
AleemMaqbool,Yes,0.961255863,0.006975077,0.649773571,-
0.065130544,0.589902975,0.077274547,0.454267737,0.107139579,0.227347322,0
.172045929,0.821201706,0.032442238,0.674534707,0.060242344,0.673389289,0.
060466932,0.589902975,0.077274547,0.720641854,0.051313278,0.42759877,0.11
3540138,0.628307242,-
0.069428382,0.629210505,0.069246338,0.720620848,0.0513173,0.620100286,0.0
71087451,0.507720977,0.09489615,0.82965329,0.030885
BahmanKalbasi,Yes,0.210629291,0.271224847,0.283541868,0.233522863,0.049111
835,0.41471336,0.008786238,0.533261751,0.004414958,0.571197109,0.20783111
3,0.272832481,0.014045553,0.504702352,0.038012847,0.435052842,0.049111835
,0.41471336,0.02407408,0.468707214,0.005706451,0.557558537,0.783181019,0.0
60710906,0.053473833,0.407698912,0.053473833,0.407698912,0.020780164,0.47
8906894,0.019602318,0.482871903,0.057711257,0.40129657
Brpkelly,Yes,0.188021973,-0.142478114,0.3153719,-
0.108896117,0.328898969,0.10591109,0.06115234,0.201591101,0.009044327,0.2

78333793,0.862719368,0.018808686,0.336568276,0.10425338,0.040919316,0.219669322,0.328898969,0.10591109,0.13132743,0.163043399,0.048054452,0.212589949,0.477462102,-0.077164255,0.332999523,0.105021748,0.056548946,0.20521512,0.045433083,0.215082788,0.017150805,0.254982411,0.036064147,0.225099484

chalexhall,Yes,0.097664454,0.171858471,0.980685987,-0.002530815,0.043577718,0.208648961,0.033205366,0.219896646,0.415233547,0.085017016,0.898189993,-0.013375217,0.022131896,0.235815577,0.056800802,0.197174077,0.043577718,0.208648961,0.067736481,0.189241143,0.208861001,0.130811549,0.214275204,0.129283184,0.037962519,0.214419271,0.287402202,0.110871231,0.202491523,0.132646176,0.015694852,0.248585876,0.395834198,0.088589351

elizabeth_news,Yes,0.411169281,-0.093150277,0.590950835,-0.060993662,0.635889006,-0.053741344,0.640734345,-0.052972602,0.974844572,-0.003581846,0.746618355,-0.036689116,0.621698225,-0.056006955,0.71973242,-0.040741444,0.635889006,-0.053741344,0.788418885,-0.030476735,0.802007506,-0.028477192,0.406369278,-0.094105168,0.676779883,-0.047325207,0.438464665,-0.087838812,0.882117089,-0.016842783,0.734992867,-0.038435424,0.737623044,-0.038039568

jglionna,Yes,0.475999759,-0.057668686,0.094556015,-0.134762934,0.238894321,-0.095154974,0.101083441,-0.132189168,0.188049053,-0.106294185,0.088777979,-0.137161662,0.134284254,-0.120812526,0.073137851,-0.144348111,0.238894321,-0.095154974,0.119158422,-0.125687431,0.365254031,-0.073218403,0.309706206,-

0.082120955,0.254883683,-0.092002542,0.108394632,-0.129457573,0.128369971,-
0.122665634,0.095226766,-0.134492085,0.055548635,-0.154108653
JimDalrympleII,Yes,0.650510714,-0.033995638,0.257078683,-
0.084911805,0.107039312,-0.120526249,0.188386918,-0.098489203,0.79298349,-
0.019696323,0.987844111,0.001143571,0.266099126,-0.083326624,0.21144772,-
0.093586711,0.107039312,-0.120526249,0.127938396,-0.113888147,0.371739704,-
0.066968416,0.204099146,-0.095105036,0.137951732,-0.111002055,0.33879273,-
0.071705004,0.18111278,-0.100126062,0.244310313,-0.087222514,0.101974725,-
0.122286167
jrosenbaum,Yes,0.801908788,0.132842233,0.801908788,-
0.132842233,0.801908788,0.132842233,0.801908788,0.132842233,0.801908788,-
0.132842233,0.801908788,-
0.132842233,0.801908788,0.132842233,0.801908788,0.132842233,0.801908788,0
.132842233,0.801908788,0.132842233,0.801908788,0.132842233,0.801908788,0.
132842233,0.801908788,0.132842233,0.801908788,0.132842233,0.801908788,0.1
32842233,0.801908788,0.132842233,0.801908788,0.132842233
lebult,Yes,0.198747452,-0.26598332,0.167021235,-0.285189233,0.103390492,-
0.333398463,0.243574202,-0.242125358,0.501208061,0.141064898,0.343809036,-
0.197572812,0.166550432,-0.285492638,0.176253013,-0.279357143,0.103390492,-
0.333398463,0.075610021,-0.361723417,0.550050214,-0.125487378,0.044642979,-
0.404941495,0.137720322,-0.305356383,0.148007313,-0.297956164,0.10296165,-
0.333789489,0.08007724,-0.356689626,0.09271992,-0.34351014
lyndsayd,Yes,0.602283132,0.094143474,0.476382105,-
0.128402922,0.985334144,0.003328216,0.82117058,-0.040909411,0.801580126,-

0.045477087,0.768784686,0.053187323,0.972268335,-0.006294192,0.341590462,-
0.170920446,0.985334144,0.003328216,0.827055534,-0.03954224,0.945607626,-
0.01235205,0.986510373,-0.003061262,0.960294338,0.009013715,0.701899316,-
0.069218529,0.536713782,-0.111506724,0.33367749,-0.173700316,0.48110558,-
0.127042361
MbasuCNN,Yes,0.493157087,0.06541888,0.774265969,0.027402312,0.476733425,0.
067920954,0.358316885,0.08761319,0.112975605,0.150598565,0.560864005,0.05
5532686,0.252231701,0.109087764,0.820405797,0.021692497,0.476733425,0.067
920954,0.814069503,0.022471655,0.60456387,0.049456968,0.437478,-
0.074098815,0.452605025,0.071683113,0.380431259,0.083671554,0.850899439,0
.017961653,0.727064218,0.033345468,0.991663023,-0.000998544
NoahGrayCNN,Yes,0.739907185,0.035682634,0.572477437,0.060628987,0.9508079
92,-0.006633053,0.85482036,-
0.019671151,0.944804399,0.00744378,0.056912699,-0.202587341,0.87662849,-
0.0166907,0.265613221,-0.119265512,0.950807992,-0.006633053,0.416809216,-
0.087136511,0.973979639,-0.003507012,0.340387286,-0.102240414,0.987077297,-
0.001741487,0.516994081,-0.069586748,0.369130407,-0.09633899,0.260761659,-
0.120474903,0.262517378,-0.120035491
phampel,Yes,0.839601748,0.036634603,0.960743917,0.008911576,0.111996823,-
0.281883795,0.182302295,-0.237991698,0.126245668,-
0.271618461,0.97140588,0.00649002,0.225197512,-0.216965103,0.141746571,-
0.261383156,0.111996823,-0.281883795,0.255446985,-0.203734995,0.259064171,-
0.20222422,0.209141097,-0.224482004,0.111996823,-0.281883795,0.158618944,-

0.251132836,0.29035319,-0.189697125,0.316500578,-0.179877446,0.184600552,-
0.236782388
PhotoBarlow,Yes,0.002227847,0.435307571,0.155440768,0.210550475,0.329621365
,0.145357422,0.811102232,0.035815617,0.016546449,-0.347953189,0.015839132,-
0.350123372,0.869788933,0.024569144,0.772330095,0.043350364,0.329621365,0
.145357422,0.701746971,-0.057359463,0.642169472,-
0.069569871,0.108069437,0.237422976,0.155203727,0.210668596,0.767805327,0
.044236271,0.605981877,-0.077206231,0.42329271,-
0.119597663,0.772330095,0.043350364
pwlabornews,Yes,0.450132228,-
0.090400215,0.223262158,0.145310728,0.576415171,-0.066932539,0.413548118,-
0.097843708,0.203033259,-0.151805078,0.986156574,0.002081262,0.645102025,-
0.055204267,0.7974938,-0.030771029,0.576415171,-0.066932539,0.570867423,-
0.067906293,0.361715239,-0.10908005,0.985326702,-0.002206041,0.647798413,-
0.054754957,0.325284886,-0.117580004,0.516300959,-0.077736835,0.570176328,-
0.068027903,0.828567145,-0.025969645
ryanfrank314,Yes,0.235491708,-0.227398151,0.537892366,-
0.119224099,0.460993705,-0.142462694,0.429246936,-
0.152639618,0.306724596,0.196589881,0.048637899,0.369328677,0.364458388,-
0.174794653,0.351916512,-0.179341335,0.460993705,-0.142462694,0.206955849,-
0.241479318,0.473935501,-0.138421199,0.08815979,-0.322309216,0.460993705,-
0.142462694,0.635242494,-0.091944008,0.206955849,-0.241479318,0.397300132,-
0.163306613,0.29438591,-0.201569557

stevegiegerich,Yes,0.573631599,-0.062270235,2.5702E-05,-
0.441936192,0.273529293,0.120843388,0.318323058,0.110200061,0.674952143,0
.046427032,0.576984147,0.06172729,0.634391923,0.052639982,0.173450031,0.1
49924762,0.273529293,0.120843388,0.707124084,0.041600134,0.869569609,-
0.018187293,0.179804203,0.147763426,0.3348153,0.106527599,0.336025002,0.1
06262795,0.684110636,0.045044488,0.281487649,0.118872458,0.347446427,0.10
3791716
thencarolsaid,Yes,0.073198673,-0.173919799,0.819288706,-
0.02234576,0.002506836,-0.28931978,0.000694549,-0.322855473,7.29727E-05,-
0.373836169,0.467980234,-0.070907221,0.005789887,-0.265089452,0.012912833,-
0.239648101,0.002506836,-0.28931978,0.017723973,-0.228881937,0.000672716,-
0.323642434,0.256271985,-0.110711034,0.002430402,-0.290176895,0.024631594,-
0.21718939,0.006714663,-0.26056819,0.002361634,-0.290969073,0.027681372,-
0.212907356
Timcast,Yes,0.97475559,0.002780739,0.687455171,-0.035341878,0.225832341,-
0.106129755,0.04815582,-0.17233798,0.024462344,-0.195782028,0.649352049,-
0.039937351,0.112127541,-0.138928742,0.195787064,-0.113312317,0.225832341,-
0.106129755,0.340338674,-0.083641195,0.050314435,-
0.170730338,0.412002368,0.071996525,0.253196513,-0.100154585,0.110478062,-
0.139564797,0.234844438,-0.104108648,0.060570476,-0.163791586,0.177157637,-
0.118175935