Uncovering geo-social semantics from the Twitter mention network: An integrated approach using spatial network smoothing and topic modeling

Advances in human dynamics research and availability of geo-referenced communication data provide an unprecedented opportunity for studying the semantics of communication and understanding the interplay between online social networks and geography. Among the most extensively studied topics in geographically-embedded communication networks, are the effect of geographic proximity on interpersonal communication; the influence of information diffusion and social networks on real-world geographic events such as group activities and demonstrations; and revealing structural and geographic characteristics of a communication network. However, little is known on how the content of interpersonal communication vary across geographic space. By integrating methods of spatial network smoothing and probabilistic topic modeling, this paper introduces an approach to extracting and visualizing geo-social semantics, i.e., how the semantics of information vary based on the geographic locations and communication ties among the users. Different from the previous work that examine the geographic variation in the content produced by individuals, the proposed approach focuses on an analysis of reciprocal conversations among individuals in a geographically embedded communication network. To demonstrate the approach, geo-located mention tweets in the U.S. from Aug. 1, 2015 to Aug. 1, 2016 were analyzed. Topics extracted from the analysis reflect geo-social dynamics of the society, way of speaking in the context of friendship, linguistic variation and the use of social media acronyms. Although the tweets were collected during primary and presidential elections, political topics discovered from the reciprocal mentions focused more on civil rights rather than the candidates and primaries. While the topic of primary candidates and elections was prominent at locations of primary elections and core supporters of candidates; civil rights was a prominent topic across the whole country.

Introduction

Advancements in mobile technology and wide use of online social networks have enabled large scale structural and geographic analysis of social ties and human communication. Previous studies utilized user generated textual communication data such as geo-referenced tweets and messages exchanged in online platforms and metadata from call detail records to study the effect of geographic proximity on social interactions (Backstrom et al., 2010; Han et al., 2017); the influence of information diffusion and social networks on real-world geographic events such as demonstrations, protests, and group activities (Vasi & Suh, 2013); and structural and geographic characteristics of the communication network (Kylasa et al., 2015; Takhteyev et al., 2012). Although such studies use information flows to model social interactions, they often are *content agnostic* - ignore the content of the information exchanged between the individuals of the network (Hansen, 1999). However, user generated content can indicate underlying interpersonal, ideological, structural and even geographic relationships between people (Lin et al., 2015).

User generated content have been the focus of researchers in information and communication sciences as well as computational linguistics. Despite the efforts that incorporate latent semantic analysis and probabilistic models to extract common topics and themes from large textual data, there has been little work (Chen et al., 2016; Kim et al., 2016) that focus on understanding of geo-social semantics of interpersonal communication, i.e., how the semantics of information vary based on the geographic locations and communication ties among individuals.

This paper introduces an approach to extracting and visualizing geo-social semantics from a geographically-embedded communication network. Different from the previous work that examine the geographic variation in the content produced by individuals, this paper presents an analysis of reciprocal conversations among individuals using an integrated approach of spatial network smoothing and topic modeling. To demonstrate the approach, over 700 million geo-located tweets in the U.S. from Aug. 1, 2015 to Aug. 1, 2016 were analyzed. First, geo-located tweets were preprocessed to extract mention tweets between personal accounts. Second, a geo-located reciprocal mention network was constructed in which a node represents an individual and a link represents the collection of mentions and replies between two individuals. Within the geo-located reciprocal mention network, each individual was assigned to an areal boundary (i.e., county) for sustaining the privacy of the user, and messages among every pair of individuals were combined into a collection of documents such as chat histories. Third, probabilistic topic modeling was performed on the collection of documents to classify each chat history into a multivariate set of topics with differing probabilities. Fourth, the individual-to-individual reciprocal mention network with classified conversations was summarized into an area-to-area network by smoothing the ego-centric network of reciprocal connections per area. Finally, topical probabilities were calculated and mapped for each area to reveal geographic and semantic patterns of communication.

Background and Related Work

In the following sub-sections, the use of Twitter as an interpersonal communication network is discussed, and a review of related work that examine the structural, geographic and semantic patterns of communication networks is provided.

Twitter as a communication network

Due to data availability and functional relationships between its users such as follow, reply, mention and retweet, Twitter has become one of the most studied communication networks. Follower, favorite and retweet functions are often used for broadcasting information and studying the process of information diffusion. On the other hand, the form of direct communication among individuals is either through private messages or replies and mentions. While a reply is a response to another user's tweet that begins with the @username of the person that she/he is replying to, a mention is a tweet that contains another user's @username anywhere in the body of the tweet. Mentions and replies allow users to join conversations on Twitter, which social interaction could be inferred as a means of direct personal communication. The context of communication can also be inferred by close observation of the textual content of the messages being exchanged.

Previous studies (Compton et al., 2014; Jurgens, 2013; Yamaguchi et al., 2013) revealed that user mentions on Twitter occur between users that are in close geographic proximity. To understand the structural characteristics, Kato et al. (2012) compared the favorite, follow and mention networks and found that all three networks are scale-free in degree distribution; and they reveal similar predominant network motifs that highlight mutual links. Cogan et al. (2012) reconstructed evolving graphs of user mentions and replies on Twitter around a particular message content and found two common typologies. The first one is "path", which illustrate back-and-forth conversations in a group of connected users. The second typology is "star", which corresponds to conversations where a single user generates a tweet to which a large number of people reply, however, the users do not respond to each other's replies. This paper focuses on the "path" typology, to study reciprocal communication among the users through the use of user mentions on Twitter.

Topic analysis

A variety of methodologies including wavelet analysis (Weng & Lee, 2011), principal component analysis (PCA) (Kondor et al., 2013), support vector machine (SVM) (2010) and generative models (Eisenstein et al., 2011) have been used to provide automatic or semi-automatic detection of relevant themes from Twitter data. Moreover, computational and semantic analysis techniques have been developed to infer human behavior, ideological and attitudinal similarity between individuals (Adamic et al., 2014), common topics and way of speaking (McCallum et al., 2007), and group identities (Tamburrini et al., 2015). Moreover, semantic analysis and probabilistic models such as Latent Dirichlet Allocation (LDA) (Chae et al., 2012; Hu et al., 2013; Hu et al., 2015; Liu et al., 2013; Pozdnoukhov & Kaiser, 2011; Zhang et al., 2009) have been successfully employed to detect geographic events, recommend places, and friends based on user location, and similarity of shared content between users in social media posts.

LDA is based on term frequency-inverse document frequency (tf-idf) (Salton & McGill, 1983), which is a statistic that takes into account the frequency of words in the corpus and reflects how important each word is to a document in a collection of documents or corpus. The tf-idf value increases proportionally to the number of times a word appears in a document. A tweet can contain up to 140 characters which do not allow multiple co-occurrences of words being used within the same tweet. Thus, training a topic model with short documents (i.e., individual tweets) results in unstable classifications with increased uncertainty due to the severe data sparsity (Yan et al., 2013). Several methods have been proposed to address the issue which suggest combining multiple tweets into document bins. Grant et al (2011) aggregated tweets into buckets based on a group of similarity measures. Hong and Davidson (2010) showed that training a topic model on aggregated tweets by users provide a higher quality and significantly better performance in classifying tweets. In addition to aggregating tweets by similarity and user, Malik et al (2013) combined tweets into bins that cover a given time range which allows the discovery of temporal changes in topics. Gerber (2014) employed space-time binning and compiled tweets into a single document based on a time window and a spatial neighborhood. Different from these studies, in this paper, tweets exchanged among a pair of individuals are combined into a document such as a chat history, and topic modeling is performed on the collection documents in order to discover themes of conversations.

Methodology

Data cleaning

The Twitter Streaming API is used to collect geo-located tweets using a geographic bounding box. Tweets with exact geographic coordinates and place names which correspond to an area (e.g., city, neighborhood) are used, while tweets with a place name at state or country level are disregarded. Geo-located tweets were preprocessed to extract mention tweets between personal accounts. The metadata provided by the API is used to filter the tweets, and users. Each tweet includes an attribute that contains whether the tweet was generated using an external application, and what that application was. A review of the contents produced by each of these applications is performed to filter tweets from non-personal user accounts such as TweetMyJobs, which is used to recruit employees, local weather reports, emergency reports, traffic crash reports, news feeds and etc. Also, tweets generated by a number of external applications (e.g., Foursquare and Instagram) are removed. Most of those tweets produced by external applications do not include conversational context. For example, Foursquare enables automatic generation of the tweet’s content with a standard text to indicate a user’s location: “I'm at Smyrna; TN in Smyrna; TN”. In addition, tweets from users with more than 3000 followers are removed to prevent any bias caused by a large number of user mentions attracted by a few users, i.e., celebrities (Lansley & Longley, 2016).

Identifying and locating pairs of reciprocal communication

One can construct an individual-to-individual communication network, where a node represents a user, and a link represents a tweet sent from user A to user B (whom user A mentions or replies to). Replies and mentions are embodied within the message of the tweet and can be downloaded using the Twitter’s streaming API. A geo-located tweet includes only the location of the sender who mentions another user (recipient). A representative location of the recipient in a mention can be derived only if the recipient has at least one geo-located tweet in the sample. In this paper, tweets from users who mention and reply each other at least once, and whose locations are known are used.

Since individuals are mobile, locations of tweets from each user are variable across space. A geo-located reciprocal mention network is constructed in which a node represents an individual and a link represents the collection of mentions and replies between two individuals. Within the geo-located reciprocal mention network, each individual is assigned to an areal boundary (i.e., county) for sustaining the privacy of the user. Tweet locations are overlayed with census data (e.g., county boundaries) to identify a home area (e.g., county) for each user based on the most frequent tweet location. Another commonly used strategy is to determine the home location based on tweets posted at night time where individuals are assumed to be home. Also, geo-located tweets have either exact geographic coordinates, or place names given in a descriptive manner such as a city name. In this paper, geo-located tweets with exact coordinates and place names that corresponded to an area at least at city scale are used.

Topic modeling of interpersonal communication

Messages among every pair of individuals are combined into a collection of documents such as chat histories. To classify the content of each chat history, a probabilistic topic model, LDA is performed. LDA provides a model of documents that assumes a collection of k topics defined as a multinomial distribution over words. In this paper, a document corresponds to a chat history which contains all the mention and reply tweets exchanged between a pair of users. This strategy allows classifying conversations rather than tweets from a user, tweets from certain time periods, or tweets from certain locations.

$$P(Z|W,D) = \frac{W_{Z+\beta w}}{total\ tokens\ in\ Z + \ \beta} * D_{Z+\propto}$$

For each possible topic Z, P (Z | W, D) is the probability that word W came from document D, which is calculated by the multiplication of $W_{Z+\beta w}$ (i.e., the frequency of W in Z), by $D_{Z+\propto}$ (i.e., the number of other words in document D that already belong to Z). β and β_w are hyper-parameters that represent the chance that word W belongs to topic Z even if it is nowhere else associated with Z (Blei et al., 2003). Based on this formula, LDA iteratively goes through the collection, word by word, and reassigns each word to a topic. Words become more common in topics where they have higher frequencies; and thus, topics become more common in documents where they occur more often. After each iteration, the model becomes more consistent as topics with specific words and documents. The model eventually reaches an equilibrium that is as consistent as the collection allows. However, it is not possible to obtain a perfectly consistent model because topics and words do not have a one-to-one relationship (Underwood, 2012). Mallet toolkit (McCallum, 2002) is used to implement the LDA model and include stop words (e.g., commonly used words such as “the”, “of”, “am”) from 28 languages prior to training the model. The topic model classifies conversations among each pair of individuals (i.e., all of the tweets between two users) with a mixture of latent topics in differing probabilities. For example, the conversations between an individual *i* and *j* might be classified as 50% about sports, 20% about fashion, 10% about food, and 10% about the other topics.

Smoothing topical probabilities over geographic areas

Once the conversations among each pair of individuals are classified into a set of topics, one can calculate the average topical probabilities per unit area. For example, among 1,000 reciprocal user pairs in Kings County NY, one can calculate the average probability of a topic such as football, by simply adding the probability of the topic per user pair, and dividing the sum by the total number of user pairs. However, because of the variable population density some counties (or areas) will have a small number of user pairs. Thus, sparse sampling of the reciprocal user pairs across small areas (i.e., the small area problem in spatially-embedded networks) result in spurious variations, where a single node or connection is often too small (with insufficient data) for deriving stable statistical measures. To address the problem, adaptive kernel smoothing can be applied to network data in order to compute and map graph measures both in space (Koylu & Guo, 2013) and space-time (Koylu et al., 2014). An adaptive kernel allows expanding the search space to include reciprocal connections of the geographic neighbors when the initial search space is found to be insufficient. This paper utilizes an adaptive kernel approach to consider connections from nearby areas. The approach is explained in the following subsections.

Neighborhood selection and kernel smoothing

Neighborhood selection is the process of determining the reciprocal connections of each area which we can define as the ego-centric network. The ego-centric network includes not only the user pairs that both users reside within the same area but also the pairs that one of the users is in the area while the other one resides in a different area are also included. A major disadvantage of an adaptive kernel approach is over-smoothing the characteristics of areas with sparse observations especially when there is an area with dense observations in the vicinity of the area with sparse observations. For example, when the ego-centric network of a rural area includes reciprocal connections from a nearby urban setting, the network measure or topical probabilities for the rural area will resemble and be dominated by that of the urban area. Also, it is likely that the content of conversations in a rural area will be different than the conversations in an urban setting.

In order to limit the influence of areas with higher density of connections on the areas with sparse connections, a similarity threshold based on connection density is used in the neighborhood selection process. The distribution of the number of user pairs for all areas is considered, and one standard deviation of gross flow per area is used as the similarity threshold. Alternatively, one can incorporate a measure of topological similarity such as one that considers the network structure (triads), or measures such as centrality, and clustering coefficient. Neighborhood selection and the adaptive kernel smoothing algorithm is introduced below.

Description of the neighborhood selection algorithm:

Definitions:

- A_i: The area i for calculating the network measure. $Ai \in A$ (the total set of n unit areas, i.e., counties).
- t: Neighborhood size threshold based on gross volume of flows.
- WF_i: The number of reciprocal pairs within A_i.
- σ: A similarity threshold to evaluate whether to include or not include a geographic neighbor into the neighborhood for smoothing. The standard deviation of WF_i. is used as the threshold.
- $N(A_i, t)$: The t-size neighborhood of an area A_i, $N(A_i, t)$, $t > 0$, is defined as the smallest $KNN(A_i, K) = \{Aj \in A \ and \sqrt{(WF_i - WF_j)\ 2} < \sigma\}$ that has a total size $\sum S_q > t$.
- $LF(A_i, t)$: The list of flows within, and in and out of the neighborhood of $N(A_i, t)$.
- $B(A_i, t)$: The bandwidth of the t-Size Neighborhood of A_i, is the radius of the smallest circle centered on A_i that covers all areas in the $N(A_i, t)$.
- K: Kernel function. Uniform function is used where all weights = 1 in the neighborhood.
- $F(A_i, t)$: $\sum_f^{LF(A_i,t)} Vol.(f) * weight\ (f)$: The weighted total volume of flows within, and in and out of the neighborhood of $N(A_i, t)$. In a kernel function (other than uniform) the weight of a flow can be calculated by the distance from the centroid of the area to the mid-point of the flow.

Steps:

(1) Compute WF, the number of reciprocal pairs within each unit area and σ, the standard deviation of number of reciprocal pairs for all units.
(2) Construct a Sort-tile-recursive (STR) tree for finding k-nearest-neighbors
(3) Determine the neighborhood

```
i. FOR each area Ai:
ii. ----IF WFi < t
iii. --------Sort the nearest neighbors of A
iv. ------------FOR each neighbor j
v. ----------------IF sqrt((WFi − WFj) 2) < σ
vi. --------------------Add j into N (Ai, t):
vii. --------------------FOR each flow in Fj
viii. ------------------------IF flow does not exist in F (Ai, t):
ix. ----------------------------Calculate flow weight based on K
x. ----------------------------Add [flow * weight] into F (Ai, t)
```

Given a positive neighborhood size threshold t based on the number of reciprocal pairs, a t-size neighborhood is derived for each area $A_i \in A$, which is the smallest k-nearest-neighbor neighborhood of A_i (including itself) that meets the size constraint.

Calculating topical probabilities per area

Given the neighborhood and the list of reciprocal pairs, $LF\ (A_i,\ t)$, the topical probabilities per area can be calculated by using the following formula:

$$P_z(A_i|\theta) = \frac{\sum_{F_{ij}\ \in\ LF\ (A_i,t)\ \ ,i\neq j}^{F\ (Ai,t)} f(i,j) * p_z(i,j)}{F\ (A_i,t)}$$

$P_z(i,j)$ is the probability of topic z in conversations among the users i and j, which at least one of them reside in the neighborhood of A_i. $LF\ (A_i,\ t)$: is the list of reciprocal pairs in $N\ (A_i,\ t)$ (i.e., the neighborhood of A), and f_A is the number of reciprocal pairs in the neighborhood $N\ (A_i,\ t)$. $P_Z(A_i|\theta)$ is the average probability of topic z given all the topical probabilities (θ) in $N\ (A_i,\ t)$.

Results

Table 1 illustrates the descriptive statistics of the geo-located tweets within the Contiguous U.S. from Aug. 1, 2015 to Aug. 1, 2016. After the data cleaning and processing, there were 2,675,130 reciprocal contacts (distinct pairs of users that exchanged tweets among each other) with 33,141,460 mention tweets exchanged between those contacts. Similar to the findings of the previous work, the amount of communication greatly decreased by increasing geographic distance. While 50 % of the geo-located reciprocal communication pairs were within the same county and 77% were within the same state.

Table 1 Tweet and user statistics

Total Tweets	700,078,319	Users	6,570,305
Tweets with mentions	221,030,872	Users > 3000 followers	249,847
Geo-located user mentions	71,438,987 (32%)	Users with tweets in only one county	1,433,870
Tweets exchanged among reciprocal contacts	33,141,460 (46%)	Users mentioned another user at least once	4,719,197
Reciprocal contacts	2,675,130	Users with reciprocal contacts	1,539,396

Topics of interpersonal communication

To evaluate the influence of parameter selection in the results of the topic model, a set of topic models were trained using 20, 50, and 100 topics with 2,000 iterations. The topical overlap among the models with different parameters were evaluated using cosine similarity. The model with 50 topics was selected based on an evaluation of overlapping topics within the model as well as the distinctness of the topics as compared to the models with 20 and 100 topics. Measures of probability (P), entropy (E) and corpus distance (CD) were used to interpret the topic modeling results. The probability of a topic represents the proportion of the corpus assigned to the topic, and calculated by the ratio of the number of word tokens assigned to the topic, to the sum of the token counts for all topics. The most interesting topics reside within the range of non-extreme values whereas extreme values indicate unreliable topics. A small probability indicates that a topic may not be reliable because we do not have enough observations to examine the topic's word distribution. On the other hand, a large probability indicates extremely frequent topic, which could be considered as a collection of corpus specific stop-words. Document entropy illustrates whether a topic is distributed evenly over conversations among many users (high entropy), or occur a lot in a smaller number of conversations (low entropy). Corpus distance measures how far a topic is from the overall distribution of words in the corpus. A greater corpus distance means the topic is more distinct; a smaller distance means that the topic is more similar to the corpus distribution.

Table 2 illustrates thirteen topics that were selected based on a probability range of 0.01 and 0.03 (the median probability of all topics). The table includes both the words and metrics of each topic. Words are ranked by their probability of occurrence from the highest to the lowest. One can infer the latent topic using the combination of the words that commonly co-occur. Some of the latent topics such as "friends & family" and "couples" do not contain words that can be used to infer the context of the conversation. These topics rather form the language elements used in conversational context such as social media acronyms (e.g., bc, ppl, ily, idk, and etc.), or words in particular dialects (e.g., yall, bruh, ima, finna, and etc.). Latent topics of "football" and "civil rights" are among the most common topics. Although the data was captured during the primary elections, mentions about primary elections and candidates is among the low probability topics. Main reason for having a low probability distribution for political topics may be that the majority of political or election related conversations are likely to be among users who do not share geographic locations of their tweets. We can also attribute the lower probability for election related mentions to the fact that most election related content are produced and retweeted within highly segregated partisan networks where there are limited connections and conversations among left and right leaning users (Conover et al., 2011; Grabowicz et al., 2012). On the other hand, user mentions in a political context often occur within a single heterogeneous cluster of users in which opposite views interact with a much higher rate than in retweet networks. However, these clusters have been observed to be less dense than more homogenous clusters of retweets (Conover et al., 2011).

Latent topics derived by the topic model are often vague in terms of the sentiment and context of conversations. This is due to the loss of sentiment and context as a result of the bag of words approach used in topic modeling. For example, words such as won, vote, voting, win, support and agree are used with any of the candidates, however, the context for their usage is lost.

Table 2 Thirteen latent topics with words and diagnostics measures

P: Probability, E: Entropy, CD: Corpus Distance

Topic	Words	P	E	CD
Football	game, team, year, win, play, football, season, qb, won, games, big, fans, teams, beat, years, week, nfl, guy, coach, defense	0.035	10.66	1.48
Civil rights	black, white, point, agree, women, isn, read, law, ppl, wrong, police, guns, problem, kids, racist, country, understand, true, cops, matter	0.030	10.54	1.65
Friends & Family	literally, bc, cute, tho, wow guys, wtf, true, ily, crying, wait, rn, idk, tweet, mom, thought, funny, ugh, honestly, bye	0.025	12.39	0.91
Couples	baby, babe, beautiful, cute, wait, birthday, amazing, bae, sweet, girlfriend, perfect, heart, boyfriend, wcw, lucky, boo, months, princess, blessed, gorgeous	0.023	11.42	1.76
Weather	snow, rain, weather, nice, cold, live, beach, storm, water, north, winter, south, week, long, beautiful, west, fun, year, weekend, hope	0.021	9.10	1.83
Faith	sis, church, jesus, twug, amen, pastor, lord, bless, faith, blessed, christ, worship, plz, family, pray, twugs, cuffmedanny, praying, sunday, word	0.021	8.27	3.22
NBA	team, game, year, lebron, win, play, curry, player, cavs, warriors, nba, kobe, won, steph, season, kd, games, finals, tho, ball	0.021	10.12	1.70
College sports	congrats, team, luck, coach, game, win, boys, season, year, big, work, job, congratulations, girls, school, awesome, ready, support, week, 2016	0.019	10.58	2.05

Baseball & Hockey	game, team, year, win, baseball, season, games, play, cubs, fans, mets, guy, hockey, guys, series, won, years trade, teams, big	0.018	10.66	1.32
Learning	team, students, work, awesome, join, excited, meeting, learning, support, event, amazing, community, check, sharing, congrats, ready, thx, job, fun, share	0.018	10.47	2.12
Primaries	trump, vote, hillary, bernie, cruz, obama, gop, president, party, clinton, voting, sanders, won, win, support, america, country, donald, candidate, agree	0.016	10.17	1.96
Driving	work, drive, money, pay, buy, ride, bike, lot, driving, nice, truck, parking, city, cars, house, park, bus, gas, street, live	0.012	11.36	1.49
Drinking	beer, drinking, cheers, photo, wine, drink, coffee, nice, ipa, bar, beers, food, awesome, dinner, enjoy, fun, tap, delicious, lunch, bottle	0.010	9.77	2.64

Table 3 represents outlier topics with high and low probabilities. “Birthday” topic has the highest probability among all topics, and represent happy birthday messages and celebrations. “Food” and “Fashion” related conversations are also quite common among the users. On the other hand, low probability topics indicate rare mentions. The two topics with the lowest probability represent mentions in languages other than English, i.e., Arabic and Spanish.

Table 3 Example outlier topics (high and low probability)

High Probability Topics				
Topic	**Words**	**P**	**E**	**CD**
Birthday	birthday, hope, pretty, bday, beautiful, amazing, hbd, ily, gorgeous, awesome, enjoy, lots, babe, sweet, aw, wonderful, fun, wait, congrats, guys	0.111 (The highest prob.)	13.04	2.37
Food	food, eat, chicken, cheese, pizza, eating, dinner, lunch, taco, breakfast, fries, sauce, hot, bacon, tacos, meat, burger, cook, wings, bread	0.061	11.19	2.50
Fashion	hair, wear, black, wearing, color, white, cute, shirt, buy, red, cut, shoes, dress, makeup, blue, pretty, nice, long, tho, pink	0.037	11.64	1.79
Low Probability Topics				
Spanish	gracias, ko, feliz, ang, batb, ng, mo, hola, amiga, saludos, dias, jajaja, nga, ay, ba, naman, quiero, noches, jajajaja, yan	0.004	7.54	2.50
Arabic	الـ لي, شي, مو, انـ ت, انـ ا, و الله, بـ س, لـ و, عـ لي, ك و, انـ ت لـ ك, الـ ي, 笑, ي, بـ و, ءاش, ـتـ و, يـ ابـ و خـ ير,	0.0006 (The lowest prob.)	7.26	5.76

Geo-social semantics of interpersonal communication

Using the adaptive kernel approach, one can produce a probability map for each topic for understanding the geo-social semantics of reciprocal mentions among users. Figure 1 illustrates the geographic distribution of two political topics: mentions of primary elections, and civil rights. Both maps in Figure 1 has the same legend which allows comparing the resulting probabilities of the two topics. Probability value refers to the commonality of the topic mentioned among individuals for the ego-centric reciprocal network of each area on the map.

The topic of primary elections consisted of candidate names, words of political context such as gop, obama, party, and candidate; and other election specific words such as vote, support, and win. On the other hand, civil rights topic was formed by commonly used words such as black, white, women, law, ppl (people), kids, police and guns, and words used in debates such as point, agree, isn (is not), read, wrong, problem, racist, country, understand, true and matter (Table 2).The word cloud represents frequently co-occurring words for each topic. Some words co-occur with a much higher frequency than others, which makes word clouds difficult to interpret. For example, the most commonly used word within the topic of primary elections was Trump, which occurred approximately three times the words Hillary and Bernie, and ten times the least frequent word agree. In order to make the word cloud more readable, font sizes are assigned based on the ranking of words within a topic. The larger the font size the highest the ranking of the word, which is assessed by its frequency within the topic.

From Figure 1.a we can infer that individuals were highly engaged in election related conversations in the North-East states of Vermont, New Hampshire and Maine; and in the north of Wisconsin and Michigan. Election related content was discussed within certain geographic locations that reflect the locations of primary elections and supporters of candidates. On the other hand, Figure 1.b highlights the metropolitan areas such as Denver, St. Louis, Washington D.C., Seattle, Portland, Minneapolis and New York City as hot spots of civil rights discussions. While the topic of primary candidates and elections highlighted localized clusters of high values in some metropolitan areas, and the North-East and rural areas in the north of Wisconsin; civil rights was a prominent topic across the whole country.

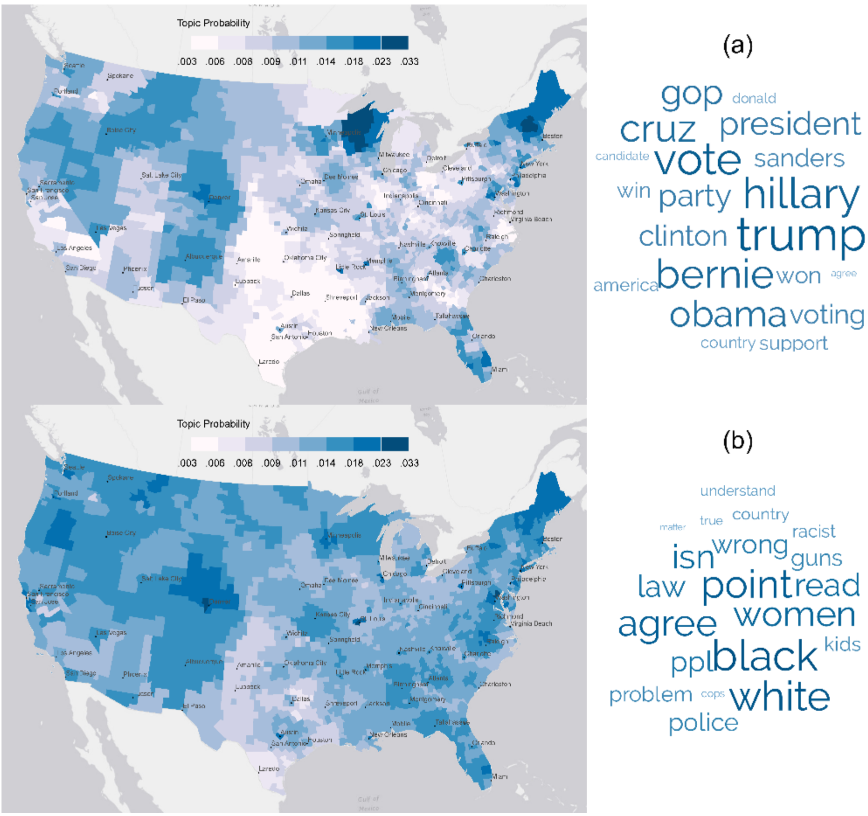

Figure 1. Topic probabilities a) Primary candidates and elections b) Civil rights. While the topic of primary candidates and elections was prominent at locations of primary elections and core supporters of candidates; civil rights was a prominent topic across the whole country.

Figure 2 highlights a clustering of “faith” topic in the South, which peaked around the states of Tennessee, South Carolina and North Carolina. Although faith is a rare topic mentioned among individuals, the clustering of high topical probabilities align well with the religious regions of the US. It is striking that coastal areas do not have as high values as the inland areas in the South. There are also regional clusters of Idaho and the north of Nevada, and New Mexico. Besides these clusters there are also spikes of metropolitan suburbs with elevated probabilities. While the words that form this topic are coherent and mostly have religious context, there is an exception of the word “cuffmedanny” which is a hashtag used in a TV series. Presence of words about this TV show in mention tweets suggest second screening (Doughty et al., 2012), which refer to individuals that live-tweet during a broadcast. In this topic, the religious references and the show were mixed in the majority of conversations.

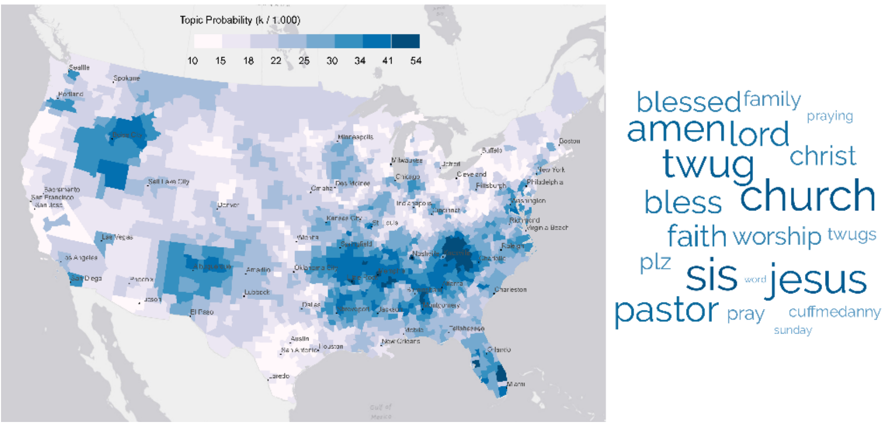

Figure 2 Topical probabilities “Faith”. Although faith is a rare topic mentioned among individuals, the clustering of high topical probabilities align well with the religious regions of the US.

Figure 3 illustrates the geographic distribution of the topic “NBA finals”. Unsurprisingly, NBA finals were predominantly discussed in metropolitan areas with major NBA teams such as Cleveland, San Francisco and Oklahoma City. Similar to the candidate names in primary elections, this topic was formed by the names of NBA teams such as Cavs and Warriors, and NBA players such as Stephen Curry, LeBron James and Kevin Durant (KD).

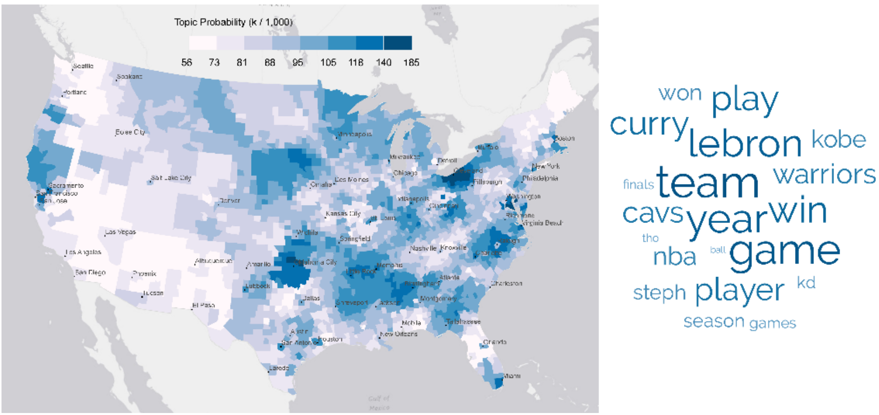

Figure 3 Topic probabilities “NBA Finals”. Unsurprisingly, NBA finals were predominantly discussed in metropolitan areas with major NBA teams such as Cleveland, San Francisco and Oklahoma City.

Discussion and Conclusion

A novel approach for extracting topical themes and their spatial patterns from a geographically-embedded interpersonal communication network was presented. The approach was demonstrated using a year of geo-located reciprocal user mentions on Twitter. The results revealed varying geographic patterns of communication on topics such as civil rights, primary elections and candidates, sports, weather, faith, food and fashion. Extracted topics reflect geo-social dynamics of the society; way of speaking in the context of friendship, and couples; and linguistic variation and the use of social media acronyms. Unlike the given time period of the dataset which covers the entire period of primary elections, mentions about the primary candidates and elections were among the least prominent topics. On the other hand, mentions about civil rights, which include race, gender and gun rights were found to be among the highest probability topics, and widely discussed across the whole country. Although the tweets were collected during primary and presidential elections, political topics discovered from the reciprocal mentions focused more on civil rights rather than the candidates and primaries. Also, individuals were highly engaged in civil rights conversations across the country, whereas election related content was discussed within certain geographic locations that reflect the locations of primary elections and supporters of candidates.

There are a number of directions for the future work. First, a major limitation of this study is that the temporal variation of the topics was ignored. The topic model can be trained to extract temporally varying topics, and the evolution of topical content over time. Unsurprisingly, the topics extracted from reciprocal mentions align well with regional geographies of semantic content such as politics, faith, and NBA. There is a need to compare the patterns of topics derived from the reciprocal communication of users with the content of the tweets generated, or retweeted without mentioning others. Such analysis would help understand the semantic variation, and the differences in geographic patterns between the interpersonal communication, and user behavior for information broadcasting on Twitter.

References

Adamic, L. A., Lento, T. M., Adar, E., & Ng, P. C. (2014). Information evolution in social networks. *arXiv preprint arXiv:1402.6792*.

Backstrom, L., Sun, E., & Marlow, C. (2010, April 26 - 30). *Find me if you can: improving geographical prediction with social and spatial proximity.* Paper presented at the Proceedings of the 19th international conference on World wide web, Raleigh, NC, USA.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research, 3*, 993-1022.

Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D. S., & Ertl, T. (2012). *Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition.* Paper presented at the Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on.

Chen, S., Yuan, X., Wang, Z., Guo, C., Liang, J., Wang, Z., . . . Zhang, J. (2016). Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. *Ieee Transactions on Visualization and Computer Graphics, 22*(1), 270-279.

Cogan, P., Andrews, M., Bradonjic, M., Kennedy, W. S., Sala, A., & Tucci, G. (2012). *Reconstruction and analysis of twitter conversation graphs.* Paper presented at the Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research.

Compton, R., Jurgens, D., & Allen, D. (2014). *Geotagging one hundred million twitter accounts with total variation minimization.* Paper presented at the Big Data (Big Data), 2014 IEEE International Conference on.

Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). *Political Polarization on Twitter.* Paper presented at the ICWSM.

Doughty, M., Rowland, D., & Lawson, S. (2012). *Who is on your sofa?: TV audience communities and second screening social networks.* Paper presented at the Proceedings of the 10th European conference on Interactive tv and video.

Eisenstein, J., Ahmed, A., & Xing, E. P. (2011). *Sparse additive generative models of text.* Paper presented at the Proceedings of the 28th International Conference on Machine Learning (ICML-11).

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. *Decision Support Systems, 61*, 115-125.

Grabowicz, P. A., Ramasco, J. J., Moro, E., Pujol, J. M., & Eguiluz, V. M. (2012). Social features of online networks: The strength of intermediary ties in online social media. *PloS one, 7*(1), e29358.

Grant, C. E., George, C. P., Jenneisch, C., & Wilson, J. N. (2011). *Online Topic Modeling for Real-Time Twitter Search.* Paper presented at the TREC.

Han, S. Y., Tsou, M.-H., & Clarke, K. C. (2017). Revisiting the death of geography in the era of Big Data: the friction of distance in cyberspace and real space. *International Journal of Digital Earth*, 1-19.

Hansen, M. T. (1999). The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative science quarterly, 44*(1), 82-111. doi:10.2307/2667032

Hong, L., & Davison, B. D. (2010). *Empirical study of topic modeling in twitter.* Paper presented at the Proceedings of the First Workshop on Social Media Analytics.

Hu, B., & Ester, M. (2013). *Spatial topic modeling in online social media for location recommendation.* Paper presented at the Proceedings of the 7th ACM conference on Recommender systems.

Hu, Y. J., Gao, S., Janowicz, K., Yu, B. L., Li, W. W., & Prasad, S. (2015). Extracting and understanding urban areas of interest using geotagged photos. *Computers Environment and Urban Systems, 54*, 240-254. doi:10.1016/j.compenvurbsys.2015.09.001

Jurgens, D. (2013). That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. *ICWSM, 13*, 273-282.

Kato, S., Koide, A., Fushimi, T., Saito, K., & Motoda, H. (2012). *Network analysis of three twitter functions: favorite, follow and mention.* Paper presented at the Pacific Rim Knowledge Acquisition Workshop.

Kim, K. S., Kojima, I., & Ogawa, H. (2016). Discovery of local topics by using latent spatio-temporal relationships in geo-social media. *International Journal of Geographical Information Science, 30*(9), 1899-1922. doi:10.1080/13658816.2016.1146956

Kondor, D., Csabai, I., Dobos, L., Szule, J., Barankai, N., Hanyecz, T., . . . Vattay, G. (2013). *Using Robust PCA to estimate regional characteristics of language use from geo-tagged Twitter messages.* Paper presented at the Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on.

Koylu, C., & Guo, D. (2013). Smoothing locational measures in spatial interaction networks. *Computers Environment and Urban Systems, 41*, 12-25. doi:10.1016/j.compenvurbsys.2013.03.001

Koylu, C., Guo, D., Kasakoff, A., & Adams, J. W. (2014). Mapping family connectedness across space and time. *Cartography and Geographic Information Science, 41*(1), 14-26.

Kylasa, S. B., Kollias, G., & Grama, A. (2015). *Social ties and checkin sites: Connections and latent structures in Location Based Social Networks.*

Paper presented at the Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015.

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems, 58*, 85-96.

Lin, Y. R., Margolin, D., & Lazer, D. (2015). Uncovering social semantics from textual traces: A theory-driven approach and evidence from public statements of US Members of Congress. *Journal of the Association for Information Science and Technology*.

Liu, X., Tang, K., Hancock, J., Han, J., Song, M., Xu, R., & Pokorny, B. (2013). A text cube approach to human, social and cultural behavior in the twitter stream. In *Social Computing, Behavioral-Cultural Modeling and Prediction* (pp. 321-330): Springer.

Malik, S., Smith, A., Hawes, T., Papadatos, P., Li, J., Dunne, C., & Shneiderman, B. (2013). *TopicFlow: visualizing topic alignment of Twitter data over time*. Paper presented at the Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara, Ontario, Canada.

McCallum, A., Wang, X., & Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research, 30*, 249-272.

McCallum, A. K. (2002). {MALLET: A Machine Learning for Language Toolkit}.

Pozdnoukhov, A., & Kaiser, C. (2011). *Space-time dynamics of topics in streaming text.* Paper presented at the Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes Twitter users: real-time event detection by social sensors.* Paper presented at the Proceedings of the 19th international conference on World wide web.

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social Networks, 34*(1), 73-81. doi:10.1016/j.socnet.2011.05.006

Tamburrini, N., Cinnirella, M., Jansen, V. A., & Bryden, J. (2015). Twitter users change word usage according to conversation-partner social identity. *Social Networks, 40*, 84-89.

Underwood, T. (2012). Topic modeling made just simple enough. Retrieved from http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Vasi, I. B., & Suh, C. S. (2013). Protest in the Internet Age: Public Attention, Social Media, and the Spread of “Occupy” Protests in the United States.

Weng, J., & Lee, B.-S. (2011). Event Detection in Twitter. *ICWSM, 11*, 401-408.

Yamaguchi, Y., Amagasa, T., & Kitagawa, H. (2013). *Landmark-based user location inference in social media.* Paper presented at the Proceedings of the first ACM conference on Online social networks.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). *A biterm topic model for short texts.* Paper presented at the Proceedings of the 22nd international conference on World Wide Web.

Zhang, D., Zhai, C., & Han, J. (2009). *Topic Cube: Topic Modeling for OLAP on Multidimensional Text Databases.* Paper presented at the SDM.