Modeling and visualizing semantic and spatio-temporal evolution of topics in interpersonal communication on Twitter

Caglar Koylu

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Check for updates

Modeling and visualizing semantic and spatio-temporal evolution of topics in interpersonal communication on Twitter

Caglar Koylu [ID]

Department of Geographical and Sustainability Sciences, University of Iowa, Iowa City, IA, USA

ABSTRACT

Interpersonal communication on online social networks has a significant impact on the society by not only diffusing information, but also forming social ties, norms, and behaviors. Knowing how the conversational discourse semantically and geographically vary over time can help uncover the changing dynamics of interpersonal ties and the digital traces of social events. This article introduces a framework for modeling and visualizing the semantic and spatio-temporal evolution of topics in a spatially embedded and time-stamped interpersonal communication network. The framework consists of (1) a topic modeling workflow for modeling topics and extracting the evolution of conversational discourse; (2) a geo-social network modeling and smoothing approach to projecting connection characteristics and semantics of communication onto geographic space and time; (3) a web-based geovisual analytics environment for exploring semantic and spatio-temporal evolution of topics in a spatially embedded and time-stamped interpersonal communication network. To demonstrate, geo-located and reciprocal user mention and reply tweets over the course of the 2016 primary and presidential elections in the United States from 1 August 2015 to 15 November 2016 were analyzed. The large portion of the topics extracted from mention tweets were related to daily life routines, human activities, and interests such as school, work, sports, dating, wearing, birthday celebration, music, food, and live-tweeting. Specific focus on the analysis of political conversations revealed that the content of conversational discourse was split between civil rights and election-related discussions of the political campaigns and candidates. These political topics exhibited major shifts in terms of content and the popularity in reaction to primaries, debates, and events throughout the study period. While civil rights discussions were more dominant and in higher intensity across the nation and throughout the whole time period, election-specific conversations resulted in temporally varying local hotspots that correlated with locations of primaries and events.

1. Introduction

The increasing availability of large-scale communication data on Location-based Social Networks has provided an unprecedented opportunity to understand the dynamics of the society, the underlying interpersonal relationships between people, and how these

CONTACT Caglar Koylu ✉ caglar-koylu@uiowa.edu

relationships vary across geography and time (Andris 2015, Lin *et al*. 2016). Extracting semantics from conversational discourse is critical for inferring human behavior, ideological, and attitudinal similarity between individuals (Adamic *et al*. 2014), common topics and way of speaking (McCallum *et al*. 2007), group identities (Tamburrini *et al*. 2015), and spatio-temporal evolution of topical themes (Pozdnoukhov and Kaiser 2011). Natural language processing methods have commonly been used to analyze georeferenced textual data to extract location-based semantics, detect geographic events, recommend places and friends based on user location, and similarity of shared content between users in social media posts (Zhang *et al*. 2009, Pozdnoukhov and Kaiser 2011, Chae *et al*. 2012, Hu and Ester 2013, Liu *et al*. 2013, Steiger *et al*. 2016). However, there has been little work (Pozdnoukhov and Kaiser 2011, Chen *et al*. 2016, Kim *et al*. 2016, Koylu 2018b) focusing on semantic evolution, and how semantics of communication vary across the social network, geographic space and time. This is due to both the lack of information on the content of interpersonal communication as a result of privacy policies and appropriate methodologies that are capable of extracting topical evolution and thematic patterns in a social network embedded in geographic space and time.

There is a critical need to analyze the spatial distribution of social relationships and extract semantics of interpersonal communication. This article introduces a framework for modeling and visualizing the semantic and spatio-temporal evolution of topics in a spatially embedded and time-stamped interpersonal communication network. There are three main contributions to this article. First, a topic modeling workflow is introduced for extracting topical themes and the evolution of conversational discourse from a time-stamped and geo-located interpersonal communication network. This framework is generic and can be applied to other social media and interpersonal communication datasets. Second, a geo-social network modeling is introduced to take into account micro-scale dyadic relationships in the modeling of aggregate-level spatial interactions. The new modeling approach enhances the previous work described by Koylu (2018b), which models macro-scale and aggregated connections (spatial interactions) among areas while disregarding the micro-level dyadic communication pairs – the connections of the social network. The geo-social network modeling approach aims to fill the knowledge gap between micro-scale social network analysis and macro-scale analysis of spatial interactions. Third, a web-based geovisual analytics environment is introduced for interactive exploration of the semantic and spatio-temporal evolution of topics in a spatially embedded and time-stamped interpersonal communication network. To demonstrate, geo-located and reciprocal user mention and reply tweets over the course of the 2016 primary and presidential elections in the United States from 1 August 2015 to 15 November 2016 were analyzed.

2. Related work

2.1. *Interpersonal communication on Twitter*

The emergence of social media emphasized the importance of online interpersonal communication and ties in diffusing information and adopting behavior in the public sphere. Through the use of its functions of retweets and mentions, Twitter enables a shared conversational context by exchanging of messages (Boyd *et al*. 2010). Retweets and mention/reply functions have been found to influence the structure of conversational discourse if it is especially related to politics (Maireder and Ausserhofer 2014).

Previous work revealed that Twitter connects diverse social networks with very different actors and modes of communication (Segerberg and Bennett 2011, Maireder and Schwarzenegger 2012).

Bruns and Moe (2013) categorize the modes of communication on Twitter into follower–followee relationships, hashtag-based exchanges, and interpersonal communication supported by user mentions and replies. Following relationship allows users to follow the updates originating from the followed user. Following relationships are often not reciprocal, because a user who is followed by another user is not required to follow that user back. Second, hashtags are used to tag tweets with a certain topic, which makes it discoverable by other users, and act as a connector among users. However, because the general purpose of the use of hashtags is to increase the visibility of tweets, it cannot be assumed that all users will follow all tweets containing the hashtag (Bruns and Moe 2013). Thus, hashtags are not suitable for inferring interaction among users. On the other hand, mentions and replies allow users to join conversations. When a user replies to another user's tweet, the @username of the person the user is replying to appear in the beginning of the tweet. On the other hand, a user can mention another user by including her/his @username anywhere in the body of the tweet. When a user mentions or replies to another user, the recipient is notified by the incoming message. Replied tweets indicate joined conversation between the users but not all mentions result in a conversation, especially when the mentioned users do not reply back. This is often the case when the mentioned account belongs to a celebrity user, brand, or institution (Bruns and Moe 2013). However, replied mention tweets indicate direct personal communication in the form of back-and-forth conversations which one can infer social interaction between users (Hong *et al*. 2011, Cogan *et al*. 2012). Previous work analyzing the structural characteristics of mention and reply tweets found that the distribution of mentions are scale-free. This indicates that there are few number of users with very large number of mentions, while most users mention others or are mentioned by others only a few times (Kato *et al*. 2012). Also, whether mentions, retweets, or following relationships, most online interactions occur between users that are in close geographic proximity (Jurgens 2013, Yamaguchi *et al*. 2013, Compton *et al*. 2014, Han *et al*. 2018).

2.2. Topic extraction and evolution

Latent Dirichlet Allocation (LDA) has been successfully employed to provide location-based recommendations, detect event locations, geographical patterns of linguistic variation, and topical themes from Twitter data (Chae *et al*. 2012, Lai *et al*. 2017, Lansley and Longley 2016, Liu *et al*. 2015, Pozdnoukhov and Kaiser 2011). LDA is based on term frequency-inverse document frequency (tf-idf), which takes into account the frequency of words in the corpus (Salton and McGill 1983). The tf-idf reflects how important each word is to a document in a collection of documents or corpus, and its value increases proportionally to the number of times a word appears in a document (Goldstone and Underwood 2012). The tf-idf measure is a 'bag of words' approach, which does not take into account the positions of the words relative to each other. The terms occurring in a large proportion of tweets are often not informative, and more specific terms occur only in a few tweets which can be captured by the tf-idf value.

Running a topic model on a corpus of tweets is challenging due to the short-text problem (Yan *et al.* 2013) which is caused by the 140 character limit. Because the limit does not allow multiple co-occurrences of words being used within the same tweet, the extracted topics exhibit increased uncertainty. To address the issue and derive robust model results, aggregation of tweets by keyword similarity (Grant *et al.* 2011), by the same user (Malik *et al.*2013), temporal binning, spatial, and space-time binning (Gerber 2014) were successfully employed. In addition to the methods of natural language processing, a variety of data mining methods such as support vector machines (Sakaki *et al.* 2010), deep convolution neural networks (Tang *et al.* 2014), wavelet analysis (Chen and Roy 2009), principal component analysis (Huang *et al.* 2016), and space-time scan statistics (Cheng and Wicks 2014) have also been used for extracting linguistic features, topics, and sentiments from geo-referenced textual data such as tweets, Yelp ratings and comments, and Flickr image tags.

As topics in a corpus with temporal information emerge, develop, and decline over time, detecting topical evolution has become an important area in text mining. Topic detection and tracking methods utilize heuristics and probabilistic models, and employ topic similarity indicators to construct temporal organization of topics (Allan 2002, Kim and Oh 2011, Zhang *et al.* 2005). A further clustering approach is applied to the topic similarity matrix to identify long-term topic chains which provide coherent topics. On the other hand, short-term or temporary topics may capture outlier themes that can be related to events. Other methods for detecting topical evolution are temporal segmentation (Gad *et al.* 2015), evolutionary clustering (Chakrabarti *et al.* 2006, Chi *et al.* 2007, Xu *et al.* 2008), and dynamic topic models (Blei and Lafferty 2006, Hu *et al.* 2015, Wang and McCallum 2006) which consider documents' timestamps as well as word co-occurrences. An in-depth discussion of the methods for detecting topical evolution can be found in Cui *et al.* (2011) and Hoque and Carenini (2016).

3. Methodology

Figure 1 illustrates the framework for modeling and visualizing the semantic and spatio-temporal evolution of topics among reciprocal communication pairs. First, the pre-processing phase is used to clean and partition the data into overlapping time windows, identify home location (area) for each user within each time window, and construct a temporal sequence of geo-located reciprocal mention networks. At this phase, mentions among each pair of users are combined into temporally organized document collections called chat histories. Topic extraction phase is then used to perform a separate LDA on the document collection of each time window. The temporally sequenced topic models produce (1) a series of word-topic matrices which represent the prominent words and their probabilities in each topic; and (2) a series of document-topic matrices which represent the prominence of topics and their probabilities in each document (reciprocal mentions) in each time window. The word-topic and document-topic matrices of the topic model series are then used to group similar topics and identify topic chains that illustrate temporally consistent topics. In the pattern exploration phase, a new geo-social network modeling approach is introduced to project the social connections and the probability of conversational topics onto geographic space and time. Finally, the topic evolution and
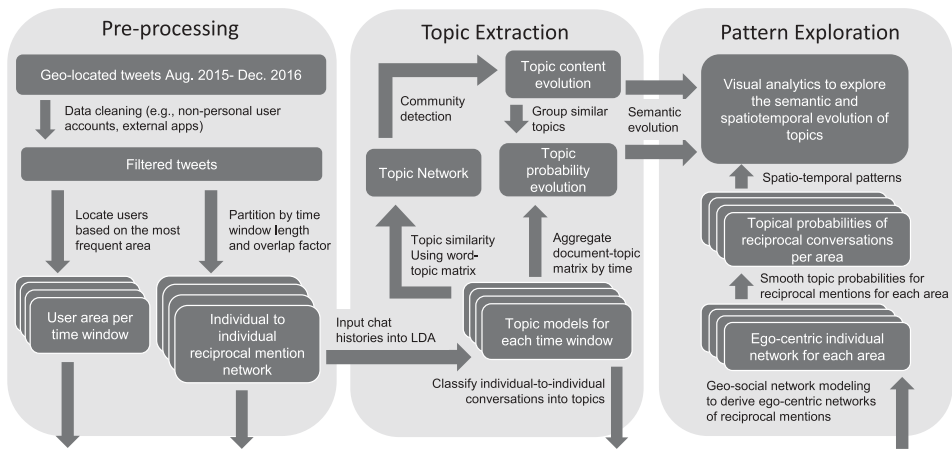
Figure 1. The framework for modeling and visualizing semantic evolution and spatio-temporal patterns of topics between reciprocal mention pairs on Twitter.

spatio-temporal patterns are linked in a web-based geovisual analytics environment to explore the changing dynamics of conversational discourse in reciprocal mentions.

3.1. Pre-processing

First, geo-located tweets within the contiguous US are collected given a geographic bounding box using the Twitter Streaming API for a given time period. Tweets from non-personal user accounts such as job advertisements, local weather reports, emergency reports, traffic crash reports, and news feeds are filtered out using the source field of the metadata provided by the API. Tweets generated by external applications are also excluded because such tweets do not include conversational context between users. For example, Foursquare application allows automatic generation of a tweet's content by tagging the location of a user: 'I'm at Iowa City; IA'. To prevent any bias caused by large volume of user mentions produced by fewer active users, e.g. celebrities (Lansley and Longley 2016), tweets from the users with more than 3000 followers are also removed at this phase.

3.1.1. Temporal partitioning and network construction

Modeling a temporally varying network as a sequence of static graphs reduces the strong fluctuations of fine-grained interactions and allows detection of general trends of evolution, thanks to temporal aggregation (Wei and Carley 2015). In this study, inter-personal communication data are partitioned into a sequence of static graphs. In order to capture the patterns in between consecutive time periods, an overlap factor of 50% is applied to monthly time windows. Overlapping (or moving) time windows help detect reciprocal links that fall in between time windows and prevent abrupt and artificial changes in the mention network and the patterns of topical and spatio-temporal evolution. In this study, the time windows start either from the 1st day of a month and expand to the 1st day of the next month, or start from the 15th day of a month and expand to the 15th day of the next month. Reciprocal communications are included in a

time window if a pair of users communicate both ways (send and receive mentions) with each other at least once within the monthly time span. R_t^w (t_{min}, t_{max}) represents a temporal sequence of reciprocal mention graphs, R_{tmin}^w, $R_{tmin+w/2}^w$, ..., $R_t^w t_{max}$, where w represents the size of a time window in a given time unit (i.e. one month). A reciprocal graph R_t^w consists of nodes defined by individuals $I = \{ i_1, \ldots, i_n \}$, and a set of reciprocal connections defined by document collection (chat histories) between individuals $D = \left\{ d_{ij}^s \right\}$, where $i \in I$, $j \in I$, $i \neq j$, $t \leq s \leq t + w$ and d_{ij}^s represents the chat history between individual i to j within the given time window.

3.1.2. *Locating users*

Geo-located tweets consist of geo-tagged tweets with exact geographic coordinates and place-tagged tweets which provide location information at the level of neighborhood, city, state, or country. If the location is provided by a place name, then the centroid of the area (e.g. the centroid of the bounding box of a city) is used to define the coordinates of the tweet. Previous work (Stefanidis *et al.* 2013) reported that the percentage of geo-tagged tweets range from 0.5% to 3%, and geo-tagged tweets increase during events such as disasters. A geo-located mention tweet includes only the location of the sender who mentions another user (recipient). Since individuals are mobile, locations of tweets from each user are variable across space, and a representative location of the recipient of a mention can be derived only if the recipient has generated at least one geo-located tweet in the sample. In this article, tweet locations are overlaid 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. Each individual is then assigned to an areal boundary (i.e. county) for both sustaining the privacy of users with exact geographic coordinates and make use of place-tagged tweets which form the majority of geo-located tweets (~87%). Instead of using the most frequent location, one can determine the home location based on tweets posted at nighttime where individuals are assumed to be home. In this article, geo-located tweets with exact coordinates and place names that correspond to an area at least at city scale are used, whereas tweets with place names at the state and country level are disregarded.

3.2. *Topic extraction and evolution*

The second phase is used to extract topical themes from chat histories defined by the edges of the reciprocal mention network of each time period. LDA is performed on the collection of chat histories $C_{R_t^w}$ to derive k number of topics within a time window R_t^w. The topic model classifies each chat history with a multivariate set of topics with differing probabilities. For example, the conversations between Robert and Brian in September might be classified as 50% about sports, 20% about video games, 10% about food, and 20% about the other topics.

Because the conversations are dynamic as new terms, phrases, and topics emerge in the conversation constantly, a separate LDA model is trained based on the collection of chat histories within each monthly overlapping time windows. LDA is a Bayesian probabilistic model of documents that assumes a collection of k topics defined as a multinomial distribution over words (Blei *et al.* 2003). Chat histories (documents) that

contain the content of the tweets exchanged between all pairs of users within a month are used to train the topic model. This strategy is based on classifying conversations rather than classifying tweets aggregated by an individual user, an area, or a time period.

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

For each possible topic Z, $P(Z|W, D)$, the probability that word W came from topic Z, is calculated by the multiplication of the normalized frequency of W in Z with the number of other words in document D that already belong to Z ($D_{Z+\ \propto}$). β and βw 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 (Goldstone and Underwood 2012). The Mallet Toolkit (McCallum 2002) is used to implement the LDA model. A separate LDA is performed on the document collection of each time window. Aggregated tweets were tokenized, and words were converted into lower case, and stop words (e.g. commonly used words such as 'the', 'of', 'am') from 28 languages are excluded prior to training the model using the Mallet Toolkit. Stemming is not applied as stemming adds noise to the results and do not improve the interpretation (Schofield and Mimno 2016).

After extracting topics from the chat histories in each time window, the next step is to identify the temporal evolution of topics. In order to detect and track the evolution of topics, a measure of topic similarity can be computed using the word-topic and document-topic matrices of the temporally sequenced topic models. In this article, pairwise similarity between topics are calculated using cosine similarity, which takes the inner product space that measures the cosine of the angle between two non-zero vectors of words that form each topic with their term (word) frequencies (Huang 2008). The cosine similarity of two topics ranges from 0 to 1 and can be described in percentages.

Using cosine similarity, topics in each time window are compared to topics both within the same window and topics across all time windows to construct a network of topic similarity. The purpose for constructing a network of topic similarity is to capture topic chains that illustrate temporally consistent topics, and their semantic evolution based on appearance, disappearance, merging, and splitting behaviors of topics. Within the network of topic similarity, a node represents a topic in time interval t, and an edge represents the cosine similarity between two topics. In order to identify similar topics, a modularity-based community detection (Clauset *et al*. 2004) is performed based on the topic similarity network. Once a topic chain is identified, measures of normalized assignment and strength evolution can be calculated using the document-topic matrices in order to evaluate how the strength and content of topics evolve over time (Hu *et al*. 2015). The normalized assignment (NA) of a topic z_k represents the total presence of the

topic throughout the reciprocal mentions in that given time window T_i, and defined as follows:

$$NA\ (z_k,\ T_i)\ =\ \frac{\sum_{j=1}^{|\ T_i|}\theta_{d_{ij}}[k]}{|\ T_i|}$$

The strength evolution (SE) of a topic z_k can then be derived by computing NA values for each time period (Hu *et al.* 2015):

$$SE(Z_K) = [\ NA\ (z_k,\ T_1),\dots,\ NA\ (z_k,\ T_t)]$$

The content of each topic is determined by the words (i.e. terms) that form each topic, and the term frequencies. The Term-Frequency (TF) illustrates the probability of a word belonging to a topic at a given-time window. A higher term-frequency indicates that the topic has a strong relationship with that term. *TF* of a word w_k in topic z_k at time window T_i is defined as (Hu *et al.* 2015):

$$TF(w_n, z_k, T_i) = \emptyset_{z_{ik}}[n]$$

The content evolution (CE) of a topic z_k is derived by computing *TF* values for words used in that topic for topic modeling results of each time window (Hu *et al.* 2015):

$$CE(z_k) = \begin{cases} TF(w_1, z_k, T_1),\dots,\ TF(w_1, z_k, T_t) \\ \qquad\qquad\vdots \\ TF(w_n, z_k, T_1),\dots,\ TF(w_n, z_k, T_t) \end{cases}$$

3.3. *The geo-social network modeling*

The topic models classify each chat history in each time period into a set of topics with differing probabilities. A classified chat history illustrates a multivariate set of topics of an undirected edge that represents communication between two individuals in a certain time window. In order to translate the topical probabilities among individual communication pairs onto a geographic area, one needs to determine the reciprocal connections associated with that area. For example, the average probability of a topic such as civil rights can be calculated for an area by simply dividing the cumulative probability of the topic by the total number of user pairs within that area. There are two major problems with such an approach.

First, considering the connections only within the area may not include the actual extent of the social network, i.e. which is defined as the ego-centric network of an area. This is because individuals have connections (i.e. friends and family), acquaintances (e.g. friends of friends, or individuals who live nearby), or online friends who may live elsewhere but could still be considered within the ego-centric network of an area.

Second, sparse sampling of the reciprocal mention pairs across small areas (i.e. the small area problem in spatially embedded networks) results in spurious variations in flow (connection) and node measures. Spatially adaptive kernel smoothing has been used to smooth flows and measures in networks embedded in space and space-time (Koylu and Guo 2013, Guo and Zhu 2014, Koylu *et al.* 2014). Using the adaptive kernel, Koylu (2018b) introduced an approach to expand the search space to include reciprocal

connections from geographic neighbors that are similar in characteristics such as con-nection or node density. However, the approach introduced in Koylu (2018b) focuses on aggregated communication flows among geographic areas and does not take into account micro-level dyadic relationships in the social network.

The effect of geographic proximity on formation of social ties has been studied widely in the literature (Backstrom *et al*. 2010, Mok *et al*. 2010, Backes *et al*. 2017). Using observational relational data through surveys and interviews, Eagle *et al*. (2009) showed strong correlation between mobility and spatial co-occurrence driven relation-ships and also highlighted the distinctness of the relationships in terms of recency and salience of interactions. Although geographic proximity plays an important role in tie formation among people, dyadic relationships that are distant apart are also essential in information diffusion, and inferring socially linked spatial communication patterns (Andris 2015). The geo-social network modeling approach introduced in this article enhances the previous method introduced in Koylu (2018b) by modeling micro-level dyadic relationships within the social network in addition to considering potential social connections from nearby geographic locations. The new approach first considers social connections (e.g. friends, family, and friends of friends) and then potential connections within nearby areas (e.g. neighbors and acquaintances).

3.3.1. *Modeling the ego-centric network of an area*

The ego-centric network of the area A_i in a time period t is defined as the list of reciprocal pairs associated with the area A_i: *LF (A_i, s) $_t$*. The algorithm to determine the ego-centric network of an *s-size area* (i.e. *the total number of links)* has two phases and is explained in Figure 2. In Figure 2(a), a black node *(e.g. a)* represents a user within the area A_i, and a solid black line represents a link between two users who are both within the same area *(e.g. a-b)*. A dark-gray node *(e.g. g)* represents a first-order network neighbor (friend), who lives outside the area A_i and has a link with a node (user) within the area (i.e. *c)*. A link between a black node and gray node is denoted as a first-order link and illustrated by a long-dashed line *(e.g. c-g)*. A light-gray node *(e.g. f)* illustrates a second-order network neighbor who is a friend of a first-order friend (i.e. *e)*. The link between a second-order friend and a first-order friend *(e.g. e-f)* is denoted as a second-order link and illustrated by a medium-dashed line.

Figure 2(a) illustrates the first phase of the algorithm. First, the size of the neighbor-hood *N (A_i, s)* is set to include *s* number of links in order to address the sparsity of connections in areas with small populations. If the total number of links is above the threshold *s*, the ego-centric network is the same as the initial network of the area, which consists of the first order links that are within and outside the area. If *N (A_i, s)* is below the threshold, the second-order links (connections with friends of first-order friends) are added to the initial network. If *N (A_i, s)* is still below the threshold, the second phase of the algorithm is performed to include connections from geographic neighbors in the ego-centric network of the area (Figure 2(b)). In the second phase, an adaptive kernel smoothing method is introduced to identify the ego-centric network for an area, and smooth the flows and network measures. Neighborhood selection process in the smoothing method is determined based on the total number of connections associated with the focal area. The ego-centric network of A_i is geographically expanded to include
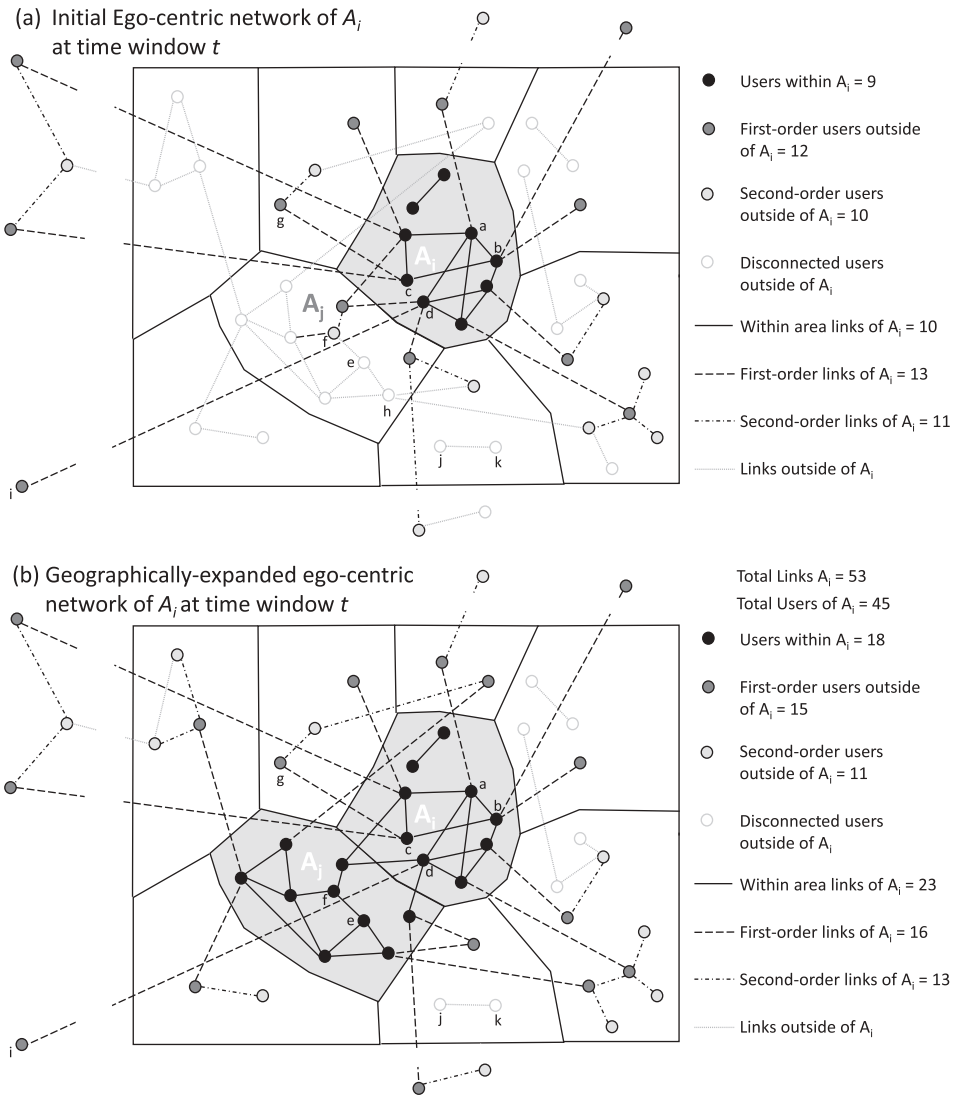
Figure 2. The geo-social network modeling and neighborhood selection.

connections from nearby areas until, $N (A_i, s)$, the total number of connections, reaches the threshold s. In this article, this threshold is set as 1000 reciprocal connections.

A major drawback of the adaptive neighborhood approach is the attribute uncertainty caused by combining connections from areas with distinct characteristics such as high density urban areas and low density rural areas. Specifically, the areas with high density of connections would dominate the areas with low-density observations if they are considered in the ego-centric network of low-density areas. Koylu (2018b) introduced a connection similarity constraint that is based on the number of reciprocal pairs within an area, and the constraint is used to take into account similarity between the characteristics of areas in the neighborhood selection process. First, the k-nearest areas of the A_i are determined based on geographic distance, and the difference between the

number of reciprocal pairs of the focal area, $N(A_i, s)$ and its neighbors is computed. If the difference is smaller than the standard deviation of the number of reciprocal pairs within all areas in time t : $\sqrt{\left(N(A_i, s) - N(A_j, s)\right)2} < \sigma$, then the k-nearest neighbor is included in the neighborhood of A_i, and its connections are added to the ego-centric network of A_i which is denoted by $LF(A_i, s)_t$.

To illustrate the neighborhood selection process and the geo-social network modeling algorithm, assume the threshold $s = 50$ for determining the ego-centric network of A_i in Figure 2(a). This threshold indicates that $N(A_i, s)$ must consists of at least 50 first-order connections within and outside the area. However, $N(A_i, s)$ is 23 in the first phase. After adding the second-order links, $N(A_i, s)$ is increased to 34, which is still below the threshold. To further expand the ego-centric network, the second phase of the algorithm is implemented to include neighbors from the k-nearest geographic areas that do not violate the connection similarity constraint (Figure 2(b)). The area A_j is the closest geographic neighbor of A_i, and the root of the squared difference between $N(A_i, s)$ and $N(A_j, s)$, i.e. the number of first order connections in A_i and A_j, is less than the standard deviation of the number of connections for all areas. Thus, the first-order connections within and outside of A_j: $LF(A_j, s)$ are added to the ego-centric network of A_i. This increases the number of connections to 39, which is still below the threshold. Next, the second-order links of A_j are added, which finally increases $N(A_i, s)$ to 52. If the threshold was not met, the next step would have been to add the connections from the next closest geographic area that satisfies the similarity constraint until the threshold is reached. In this article, the number of connections is used for defining the threshold; however, one can use the number of users or structural characteristics such as number of triangles or cliques in the ego-centric network.

3.3.2. *Smoothing topical probabilities in space-time*

After selecting the ego-centric network, a kernel function can be incorporated to weigh each flow (connection) within the network based on the distance of nodes to the focal area. This process would weigh mention pairs where both users are within the area more than the pairs that only one user is in the area, or none of the users are in the area (i.e. in the context of second-order links, the users may not be within the focal unit area). The most commonly used kernel functions include the uniform kernel, the Gaussian kernel, and triangular (linear) kernel. In this article, the uniform kernel, which assigns the same weight to each connection, is used. This choice is optimal within the context of online communication since the kernel smoothing with a distance decay function such as the Gaussian or the triangular kernel may result in undesired impact of diminishing the influence of first- and second-order links outside the focal areas.

Given the ego-centric network of an area A_i in time window t: $LF(A_i^t, s)$, the weighted average of the topical probabilities is calculated using the following formula.

$$P_z\left(A_i^t|\theta\right) = \frac{\sum_{i \in LF(A_i, s)^t}^{N(A_i, s)^t} \sum_{j \in LF(A_i, s)^t, i \neq j}^{N(A_i, s)^t} w_{A_i}(i, j)^t * p_z(i, j)^t}{N(A_i, s)^t}$$

$P_z\left(A_i^t|\theta\right)$ is the average probability of topic z given the probability of all topics (θ) in the neighborhood of A_i: $N\left(A_i, s\right)$ within the time interval t. $p_z(i, j)^t$ is the probability of topic z in conversations among the users i and j in time window t. $w_{A_i}(i, j)^t$ is the weight of the connection determined by a kernel function, and the distances of the individuals i and j to the area of the area A_i. The uniform kernel function is used in this article, and the weigh equals to 1 for all connections. $N\left(A_i, s\right)^t$ is the total number of reciprocal pairs in the neighborhood of A_i at time t. $LF\left(A_i^t, s\right)$ is the list of flows in the s-size network of A_i within the time interval t.

4. Results

Geo-located tweets in the contiguous US between 1 August 2015 and 15 November 2016 were collected using the Twitter Streaming API. The percentage of the tweets with exact coordinates was 12.4% in this dataset. Tweets with place names that were at the state or country level were excluded, which corresponded to 15% of place-tagged tweets. After the initial filtering process, the dataset consisted of 898,182,251 tweets and 7,321,492 users. The number of users with more than 3000 followers was 217,658. The number of tweets that included a user mention was 320 million (35% of all tweets), and 4.8 million users were mentioned at least once. The monthly frequency of reciprocal mentions had intense fluctuations in October and December 2015, and March 2016, and there was an overall increasing trend toward the end of election period with an average of approximately 225 thousand reciprocal communication pairs per month (Figure 3).

4.1. *Parameter selection for the topic model series*

This section presents the criteria and experiment results for parameter selection in the topic modeling workflow. In order to determine the optimum number of topics that minimize the number of duplicate topics, an evaluation was performed using topic models with 25, 50, and 100 topics with 2000 iterations for each of the overlapping time windows. In a topic model with fewer number of topics, unique topics are often combined into fewer topics, whereas larger number of topics result in topics with overlapping content (Newman *et al*. 2010). In order to find the optimum number of
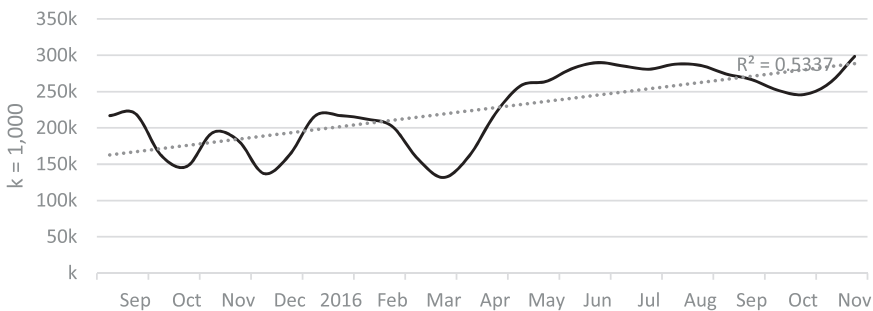
Figure 3. The number of reciprocal mention pairs between 1 August 2015 and 15 November 2016 using a one-month moving average window.

topics across all time periods, a topic model with each k number of topics (i.e. 25, 50, and 100) were compared to one another consecutively for each time window. Cosine similarity was used to measure the similarity of topics per time window. The following criteria was used to determine unique topics between two model outputs: (1) the topic must have a probability threshold above 1% in order to eliminate noisy topics that only exist in a small number of documents. (2) The topic in the model with larger number of topics should have a cosine similarity less than 20% with any of the topics of the model with smaller number of *k* topics. Following this criteria, the 100-topic models were compared with the 50-topic models, and the 50-topic models were compared with the 25-topic models for each time window. The 100-topic models introduced an average of 25 overlapping topics within the same model, and an average of 1.5 unique topics as compared to the 50-topic models in all time periods. On the other hand, the 50-topic model introduced an average of nine overlapping topics within the same model while introducing an average of 1.87 unique topics as compared to the 25-topic models in all time periods. Finally, the 25-topic model introduced an average of three overlapping topics within the same model. The models with $k = 50$ number of topics were selected as it produced less overlapping topics than 100, and more distinct topics than 25.

4.2. *Topic chains, content, and strength evolution*

A separate topic model was used to train chat histories as documents for each of the 30 overlapping time periods. Fifty topics were extracted from each of the monthly collection of chat histories, which generated a total of 1500 topics. Cosine similarity between the word frequency vectors of each topic in all time periods was computed, and a network of topic similarity was constructed to link topics within the same time period and topics from different time periods. Within the network of topic similarity, a node represents a topic in a time interval, and an edge represents a binary link that indicates whether a pair of topics can be considered as similar. A threshold of 80% cosine similarity was used to determine if two topics are similar. As a result, the topic similarity network consisted of 1,500 topics, and 6,746 edges that illustrate the number of topic pairs that are similar to each other at a rate greater than 80%.

A modularity-based community detection algorithm (Clauset *et al*. 2004) was performed on the topic similarity network in order to group similar topics, extract topic chains, and merging and splitting topics. Large communities detected by the community detection algorithm represent topic chains that connect similar topics in a temporal sequence. Topic chains that are formed by topics in a time series are often semantically meaningful, whereas singleton topics are less coherent (Kim and Oh 2011). The top 16 largest topic chains with the top 20 words with the highest cumulative term frequencies across all time periods are reported in Table 1. Labels of the latent topics can be inferred using the combination of the words, and observing the original reciprocal mention tweets classified by each topic. Among the temporally consistent topic chains are work and school activities, sports fandom (Bruns *et al*. 2013); daily routines such as watching, drinking, wearing, music and eating (Java *et al*. 2007), and election and civil rights topics that are related to the given time-period (Table 1).

In order to demonstrate the utility of the proposed methodology, this article focuses on the two topic chains: civil rights and elections. Figure 4 illustrates the content

Table 1. Temporally consistent topic chains of reciprocal communication. The words with the highest term frequencies in all time periods were selected to illustrate each topic.

Topic	Words
Civil rights	Black, point, white, agree, wrong, true, understand, read, lot, women, ppl, person, real, problem, police, racist, matter, cops, opinion, thought
Elections	Trump, vote, Hillary, Bernie, Obama, president, GOP, cruz, support, Clinton, voting, party, America, country, win, Donald, candidate, Sanders, HRC, election
School	School, class, year, high, college, classes, math, test, grade, teacher, study, semester, senior, freshman, English, week, luck, paper, middle, kids
Sports	Game, team, year, win, play, baseball, season, coach, football, fans, games, cubs, playing, player, soccer, congrats, beat, won, luck, big
Birthday	Birthday, hope, bday, pretty, ily, beautiful, amazing, hbd, babe, gorgeous, enjoy, awesome, lots, sweet, wonderful, friend, thx, baby, boo, wait
Watching	Watch, movie, watching, season, episode, watched, movies, netflix, favorite, star, wait, episodes, read, story, film, book, started, series, finished, loved
Drinking	Beer, drink, cheers, drinking, wine, water, coffee, bottle, drunk, nice, bar, beers, shots, awesome, alcohol, enjoy, drinks, tea, ice, party
Dating	Baby, cute, babe, ily, avi, beautiful, wait, boo, guys, friend, hot, sweet, wow, luv, literally, perfect, bae, pretty, heart, friends
Bro	Bro, y'all, bruh, tho, yea, bout, boy, hell, fam, real, smh, ima, lit, lil, tryna, chill, yall, way, work, wit
Wearing	Hair, wear, wearing, shirt, black, white, color, shoes, red, cut, blue, dress, buy, nice, pants, cute, long, clothes, pair, makeup
Music	Song, music, listen, album, songs, listening, drake, heard, band, favorite, video, hear, play, sing, check, rap, playing, lyrics, live, kanye
Video games	Play, game, playing, games, team, xbox, played, guys, join, stream, cod, win, link, fun, skype, ill, beat, add, watch, send
Money	Money, pay, buy, work, free, tickets, job, ticket, paid, card, bought, check, store, sell, lot, worth, paying, working, tax, price
Food	Eat, food, pizza, chicken, cheese, eating, bring, ice, cream, hot, ate, dinner, chocolate, taco, breakfast, cake, fries, milk, coffee, sauce
Work	Work, wait, weekend, week, coming, guys, friday, excited, party, saturday, fun, ready, year, visit, sunday, y'all, break, trip, summer, friends
Bro birthday	Bro, birthday, big, brother, dude, fam, homie, boy, congrats, lil, brotha, hope, bday, buddy, bud, dawg, preciate, luck, guy, stay

evolution of the two topics in six monthly time periods. From this figure, one can identify the content evolution of topics by tracing the appearance of new terms, disappearance, the changing order and prevalence of the terms. These monthly periods were selected based on the most distinct shift of content, and the merging events of election and civil rights topics. A topic is considered as a merged topic (i.e. purple) when its cosine similarity is equal to or above 80% for both civil rights and election topics from the previous and next time periods. The ordering of the words in term tables indicates the changing importance (frequency) of terms in election topic over time, and each time period is labelled by the major events associated with conversations in each time period. These events were Iran Deal (August 2015), the terrorist attacks in Paris and San Bernardino (December 2015), primary elections (March 2016), black killings (July 2016), Presidential Debates (October 2016), and Presidential Elections (November 2016). The civil rights and election conversations existed separately or were merged during time periods when significant events happened. The last time period prior toward the presidential elections from 15 October to 15 November 2016 generated three separate topics about the elections and civil rights. The first topic was a merged topic of election and civil rights, the second topic was about voting participation, and Electoral College, and the third topic was about Hillary Clinton's email controversy. A sample of tweets
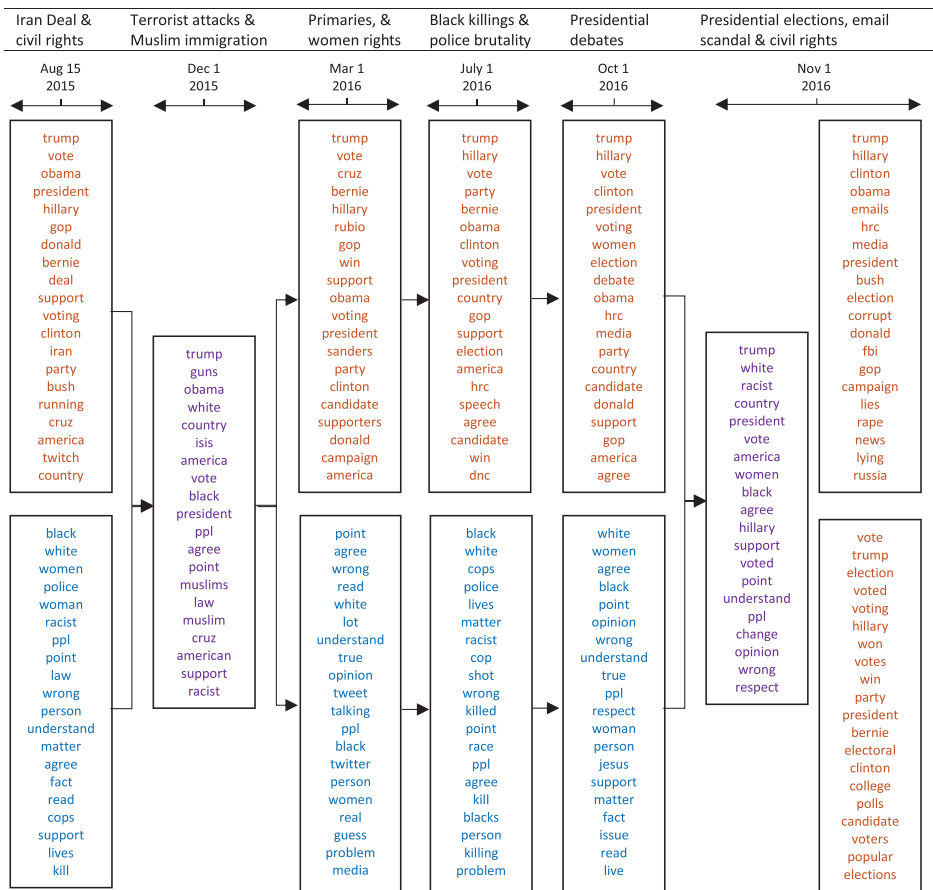
Figure 4. Content evolution of **election** and **civil rights** topics in relation to events in the news media using the top 20 words that form each topic. The six time periods were selected to illustrate the merging of the two topics and the largest shifts in the content evolution of the two topics. The two topics merged in 15 November 2015–15 December 2015 and 15 October 2016–15 November 2016 time periods.

classified by these three topics in this time period are anonymized and listed in Appendix (Table A1) to provide accuracy and context for the topic labels.

Figure 5 illustrates the content and strength evolution, and popular term frequencies of 'election' and 'civil rights' topics. The term Trump was used more than four times than Clinton or any other candidates for the majority of the whole time period. This finding agrees with and amplifies Gamio and Brochers (2016)'s analysis results on the Goggle News homepage which revealed that Trump appeared on the news twice as more than Clinton did. Unsurprisingly, fluctuations in the intensity of term frequencies for candidates were associated with the events in news media about the primaries and candidate campaigns. For example, conversations about the candidate, Ted Cruz, declined soon after he suspended his campaign for presidency. Similarly, the mentions about Bernie Sanders started to decline right before Hillary Clinton was nominated for presidency by the Democratic Party (Figure 5(a)). In addition, the civil rights topic exhibited high
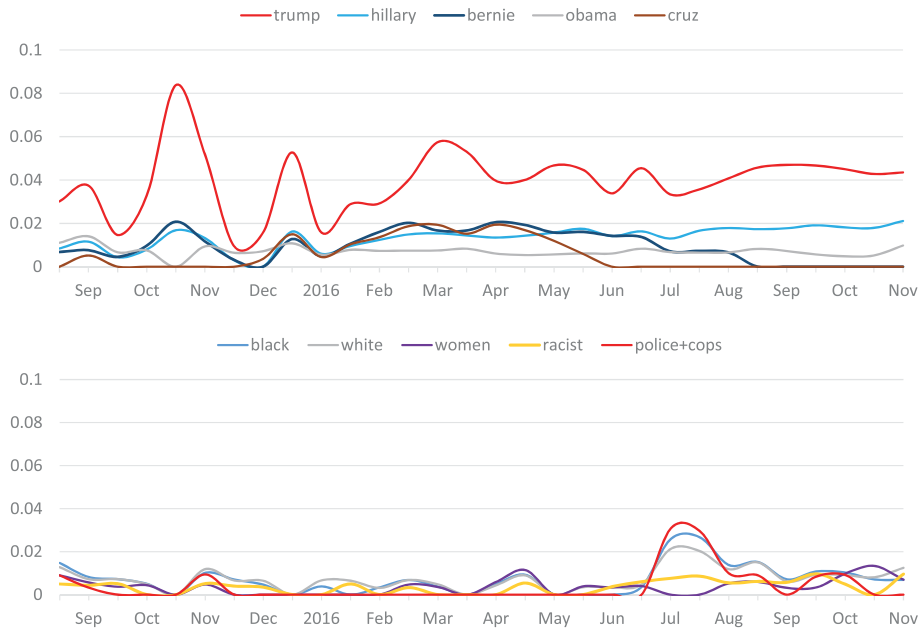
Figure 5. Probability of the top five contextual terms in (a) election topic: trump, hillary, bernie, obama and cruz. (b) civil rights topic: black, white, women, racist, police, and cops.

frequencies for the terms police, cop, black and white between July and August. 2016 as a reaction to the police killings of two black individuals, and the statements of the candidates about this event (Figure 5(b)).

4.3. Spatio-temporal and semantic patterns

Spatio-temporal and semantic patterns can be explored using the web-based geovisual analytics environment at the link: http://www.geo-social.com/topicviz/mentiontopics. html. In order to capture the context related to the study period related to elections, the visual analytics environment uses election and civil rights topic chains with their semantic and spatio-temporal evolution. However, this environment can easily be configured to visualize other topics and topic chains, as well. A snapshot of the environment is illustrated in Figure 6. The environment is composed of three coordinated panels: a temporal panel, a topic panel, and a map panel. The temporal panel consists of a bar chart for depicting the strength evolution of the two topics over each of the overlapping time intervals, and a time slider for changing the time interval. The bar chart illustrates the strength evolution (popularity) of civil rights (blue) and election (orange) topics over time, and the merging of these two topics throughout the entire period. The topic panel includes a word cloud of commonly co-occurring words that define each topic within the selected time interval, and a drop-down list that allows switching between the topics for map display if there are more than one topic within the selected time interval. Finally, the map panel includes a county-level choropleth map illustrating the probability of the selected topic within a selected time interval. The user can move the time slider to switch between the time intervals and visualize the topics within each time
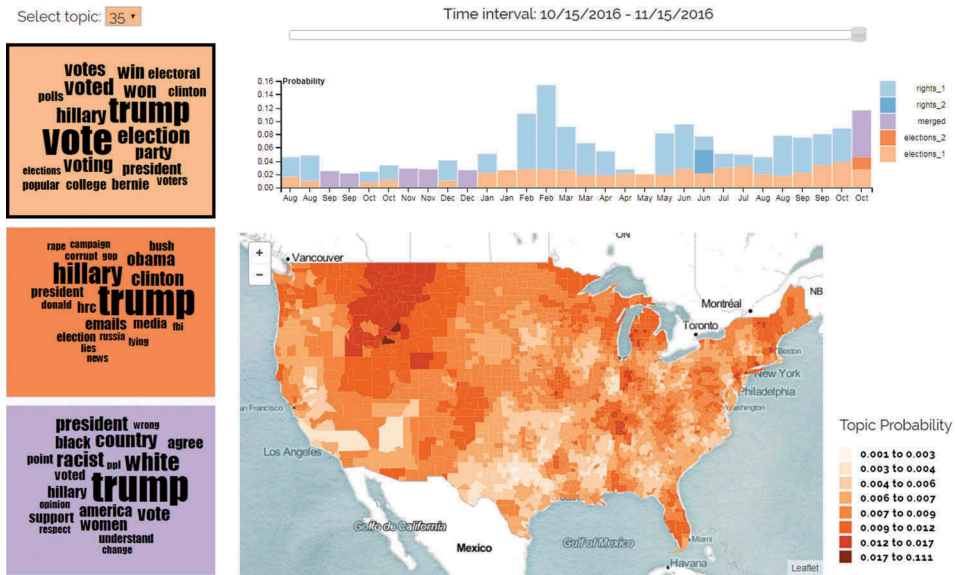
Figure 6. The geovisual analytics environment for exploring semantic and spatio-temporal evolution of election and civil rights topics on Twitter.

interval. The map panel is updated to illustrate the spatial patterns of a selected topic using the dropdown list. The three color hues are used to link the three panels. Orange illustrates election-related topic, blue illustrates civil rights topics, and purple illustrates merged topic. Notice that there could be more than one topic from the same category (e.g. two election related topics). In the topic panel, while color hue is used to distinguish civil rights, election, and merged topics, color brightness is used to distinguish topics within the same category. The topic probabilities of civil rights, election, and merged topics in all time periods for each county were combined onto the same number line, and a quantile classification was applied to derive eight classes. The choropleth map in Figure 6 illustrates the spatial distribution of the election topic 35 in the last time period 15 October–15 November 2016. The topic 35 is selected in the dropdown list and marked with a thick outline around the topic word cloud in the topic panel.

In order to summarize the findings of the pattern exploration phase, small multiples of the choropleth time series for the civil rights and election topics are provided in Figure 7. The first period was 1 February–1 March 2016, which represented a significant increase in the probability of civil rights topic (Figure 6) due to the starting of the primary elections. The election topic (Figure 7(a)) highlighted local hotspots in the upper North East, New Mexico, Colorado, Florida, Tennessee, the coastal areas in the Carolinas, and suburban areas surrounding big cities such as Minneapolis, Chicago, and New York City. Most of these areas were the places that primaries were held in February 2016 and the following time period in March 2016. Civil rights resulted in higher intensity of conversations across the Nation with a focus on racial and gender related conversations (Figure 7(b)). The second period was 1 July–1 August 2016, which marked the end of
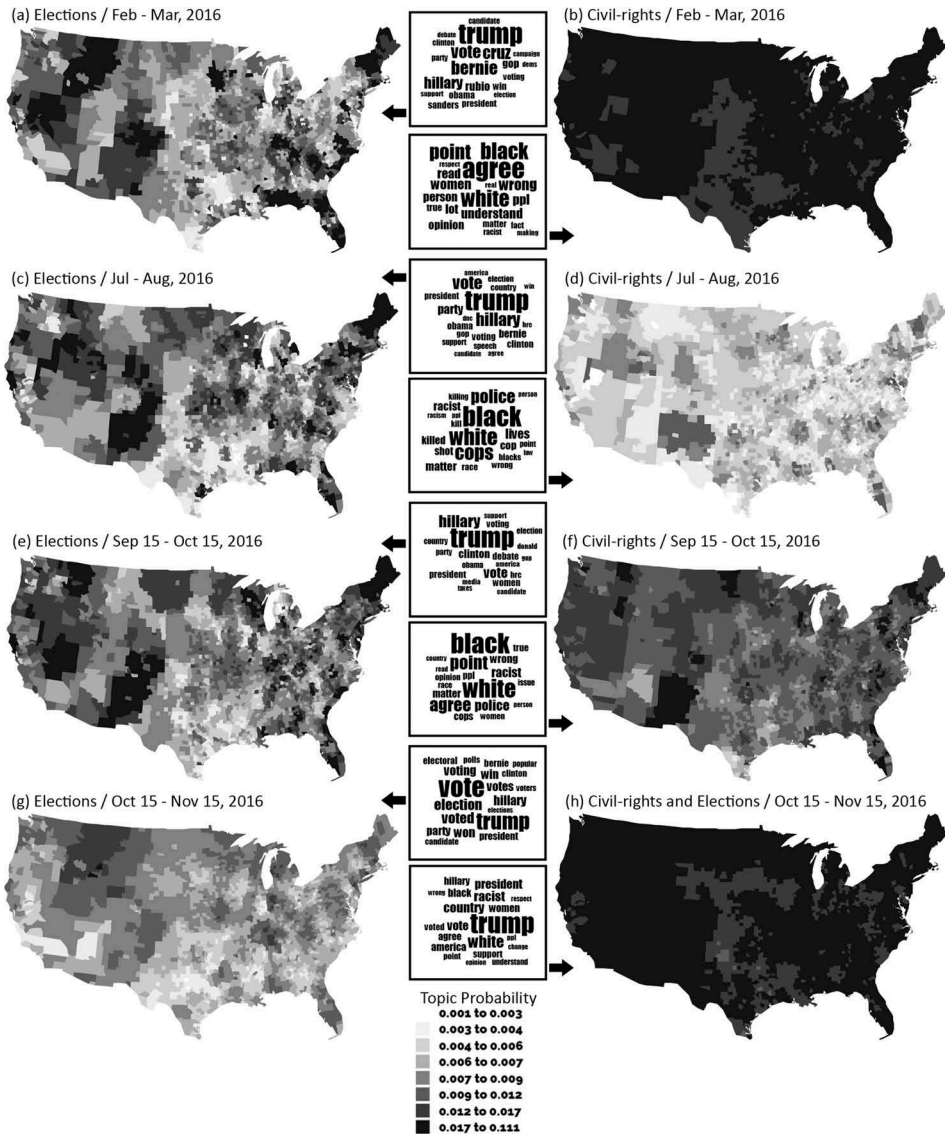
Figure 7. Spatio-temporal and semantic evolution of election and civil rights topics.

Democratic primaries. Distinct shifts in the content evolution of the election and civil rights topics were observed in this time period (Figure 7(c–d)). The terms Trump, Hillary, and Bernie remained to be dominant, while the other candidates did not appear in the top co-occurring words in election discussions. Many of the hotspots of the election topic from earlier time periods remained to be similar in geographic distribution such as the upper North East, New Mexico, and Colorado, and surrounding areas of the large metropolitan cities such as Chicago and New York City. As compared to other time periods, content of the civil rights topic was related to black-lives-matter movement and the conversations ignited by the black killings happened in early July (Figure 7(d)). Although the topical content was specific to the discussions of racial bias, the civil rights

topic in this time period was lower in intensity, and locally varying in its geographic distribution. The third period was 15 September–15 October 2016 (Figure 7(e–f)). The election topic was primarily about the debates happened in September 26 and October 9, while the civil rights topic included gender rights back in the top co-occurring words. Both temporal and geographic intensity of the civil rights topic was higher as compared to the previous period in July–August 2016.

Finally, the fourth period was 15 October–15 November 2016 (Figure 7(g–h)), which included the conversations prior to and a week after the Presidential Elections on 8 November 2016. There were three interrelated topics about elections and civil rights (Figure 6). The first topic in this period was about voting behavior and activity. The second topic was about Hillary Clinton's email controversy and the FBI investigation. The third topic was the merged topic of election results and civil rights debates among people. Figure 7(g–h) illustrates the geographic and semantic patterns of the first and the third topic. Overall, the intensity of the election topic was less than one-third of the merged topic in this time period, and the election topic was primarily about users' share of voting activity (Figure 7(g)). The third topic was the merged topic of civil rights and election during this period (Figure 7(h)), which had the highest intensity (probability) across all time periods. The topic included the candidates Trump and Hillary within the top 20 words alongside with the civil rights-related terms such as black, white, and women; voting terms such as voted and vote; and the terms such as racist and respect. In comparison with the probability of election and civil rights topics in all time periods, this topic was widely discussed across the country with higher intensity (Figure 7(h)).

4.4. *Topics of reciprocal mentions versus user-aggregated tweets*

To identify similarities and differences between the content generated by individual users and the content of reciprocal mentions, tweets were aggregated by users into documents for each of the overlapping time windows, and topic models were trained using the user-aggregated document collection per time period. Because there is no justifiable reasoning for removing of mentions from user-generated content, both mentions and replies were kept in user-aggregated documents. An empirical evaluation was then performed to answer the following questions: Are there specific topics covered in reciprocal mentions but not in user-aggregated tweets and vice versa?

The topics derived from user-aggregated tweets were compared to topics of reciprocal mentions in terms of the distinct topics between the two models, and the average percentage of overlapping topics within each model. Distinctness was measured by the number of topics that had similarity greater than or equal to 80% between the user and mention model for each of the time period. On the other hand, within topic overlap was measured based on the average pairwise similarity between all topics within each model across the whole time period. Overall, user topics generated more distinct topics than the mention topics (Figure 8(a)); however, user topic models resulted in higher overlap within the same model, while mention topics produced less overlapping topics across the whole time period (Figure 8(b)). In addition to distinctness and within-model overlapping, there is also strong overlapping of topics among the user and mention models throughout the whole time period. A significant part of this overlap can be explained by the inclusion of mention tweets in the user-aggregated documents. Otherwise, exclusion
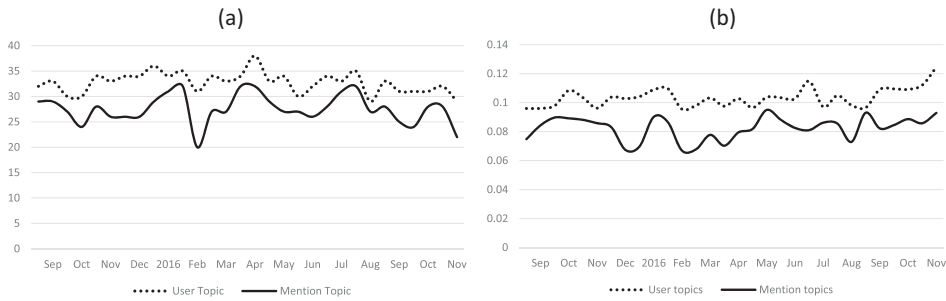
Figure 8. Comparison of mention topics with user topics. (a) The number of distinct mention and user topics. (b) Average within-model overlap ratio for mention and user topics.

of tweets that include mentions would result in bias in terms of the true representation of each individual's posts.

In order to highlight the differences between the two topic models, Table 2 illustrates the comparison of mention and user topics for the last time period which covers the presidential elections: 15 October–15 November 2016. Topic modeling by user-aggregated documents generated 7 topics related to election and civil rights. These topics were strongly overlapping with each other in this time period. The two distinct user-topics are related to news sharing (topic 36 in Table 2) and second screening or live-tweeting about the elections (topic 48 in Table 2). These two topics have the least similarity with the mention topics as well as with the rest of the user topics. While the minimum similarity among the topics 0, 3, 24, 28, and 40 was 76% for the topics 24 and

Table 2. Comparing user and mention topics related to elections and civil rights.

ID	Topic words	Prob.	Target ID	% Overlap
	User topics compared to mention topics			
0	Trump, Donald, women, GOP, Trump's, president, white, Hillary, rigged, racist, woman, America, vote, Putin, sexual, election, lies, debate, campaign, nasty	0.01562	49	0.82
3	Trump, Clinton, Hillary, election, Obama, campaign, video, Wikileaks, FBI, news, emails, Russia, media, law, Donald, debate, fraud, voter, black, police	0.01384	49	0.82
24	Trump, Hillary, vote, Obama, media, America, Clinton, country, HRC, corrupt, news, American, president, truth, corruption, lies, CNN, Americans, liberal, job	0.01983	49	0.84
28	Trump, white, president, country, America, vote, racist, voted, election, women, hillary, black, years, support, Obama, Donald, voting, supporters, American, respect	0.0756	10	0.90
36	Trump, election, party, years, vote, point, read, media, political, candidate, voters, white, agree, work, wrong, GOP, lot, left, real, true	0.04967	35	0.49
40	Trump, vote, president, Hillary, election, Donald, America, Clinton, voted, win, won, voting, country, wins, Obama, votes, electoral, college, Florida, years	0.07471	35	0.81
48	Election, vote, Trump, America, women, watching, read, woman, debate, news, voted, work, years, watch, white, live, country, voting, hope, president	0.05323	35	0.54
	Mention topics compared to user topics			
10	Trump, white, racist, country, president, vote, America, women, black, agree, Hillary, support, voted, point, understand, ppl, change, opinion, wrong, respect	0.07051	28	0.9
35	Vote, Trump, election, voted, voting, Hillary, won, votes, win, party, president, Bernie, electoral, Clinton, college, polls, candidate, voters, popular, elections	0.02699	40	0.81
49	Trump, Hillary, Clinton, Obama, emails, HRC, media, president, Bush, election, corrupt, Donald, FBI, GOP, campaign, lies, rape, news, lying, Russia	0.01873	24	0.84

40; the topic 36 was most similar to topic 48 with 52%, and the topic 48 was most similar to topic 28 with 61%. As stated earlier, the fact that the mention topics have strong overlap with user topics can be explained by the presence of reciprocal conversations within the user topic models.

Overall, user-topics were generally formed by personal views with partisan content, mention topics included debates, arguments, and words used in conversational discourse. The results of the comparison revealed that the majority of civil rights and election topics were discussions and debates among people with opposite views. Sample tweets provided in Appendix (Table A1) illustrate that these opposing views were not always between Democrats and Republicans, but also between Democrats with different views (e.g. the supporters of Bernie Sanders and Hillary Clinton).

5. Discussion and conclusion

Extracting the semantic evolution and spatio-temporal patterns from an interpersonal communication network is critical for a comprehensive understanding of the variation in conversational discourse, the changing dynamics of interpersonal ties, and the digital traces of social events. Besides the privacy issues and unavailability of data, it is challenging to study interpersonal communication since a comprehensive analysis requires simultaneous consideration of the social network connections, semantic and spatio-temporal variation in communication among individuals. From a methodological perspective, this article introduced (1) a workflow for modeling topics and extracting the evolution of conversational discourse in an interpersonal communication network; (2) a new network modeling approach to project connection characteristics and semantics of communication onto geographic space and time; (3) a web-based geovisual analytics environment for extracting semantic and spatio-temporal evolution of topics in a spatially embedded and time-stamped interpersonal communication network. The proposed framework is generic and can be applied to a variety of human communication datasets. From a thematic perspective, the findings of this article contributes to the analysis and evolution of topical themes during the period of 2016 primary and presidential elections in the United States. The results revealed that the large portion of the topics extracted from mention tweets were related to daily life routines, human activities, and interests such as school, work, sports, dating, wearing, birthday celebration, music, food, and second screening. On the other hand, political conversations were split into civil rights and election topics. Civil rights and election topics were also temporally consistent; however, both topics evolved significantly in terms of their content and intensity over time. Hotspots of where these conversations took place changed based on events such as political debates, primaries, police killings, as well as the election results. Our results confirm the findings of the previous work such that Twitter functions both as a social network that is geographically close with each other, and as a news distribution network in which individuals are geographically far apart from each other (McGee *et al*. 2011, Koylu 2018a).

The comparison of topics generated by user-aggregated tweets versus the topics of reciprocal mentions showed both distinct and overlapping content between the two models. User topics generated more distinct topics, however, within-model overlapping was higher in user topics than the mention topics. In addition, the two topic models

illustrated strong overlapping content, which can be attributed to the inclusion of mention and replies in user-aggregated tweets for training the user topic models.

There are a number of limitations to this study. First, demographics of twitter users are heavily biased toward younger age groups and cannot be reflective of the general population. Moreover, mention and reply tweets represent only a small portion of inter-personal communication which mostly happen in person, through phone calls, text messaging, and video conferencing. However, the approach introduced in this article can readily be applied to any form of communication data with spatial information in order to discover semantic and spatio-temporal evolution of conversational discourse without revealing privacy of individuals. Also, geo-located mention and reply tweets used in the case study represent only a sample of all mention tweets. Constrained by the opt-in behavior of users for geographic location, the dataset does not include all mention tweets. Future work is needed to increase the completeness of the mention network for a more comprehensive analysis of the conversational discourse on Twitter.

This article introduced an analysis of a monthly resolution of reciprocal communica-tion due to the sparsity of geo-located tweets. There is a need to develop a multi-scale resolution for capturing the evolution of topics and spatio-temporal patterns at different levels of space-time granularity. This would allow detecting hourly, daily, weekly, sea-sonal, and cyclical patterns of communication, and better understand the variability of conversational discourse. Another potential direction for future work is to differentiate the topical context by taking into account the network structures in the online com-munity (Smith *et al.* 2014). Different mechanisms and processes are behind spatio-temporal patterns of topics. For example, an increased intensity of a topic in a geo-graphic area may indicate the local community's response to a local event such as a primary election. However, the intensity of a topic probability could also be caused by users who are outside the area with strong connections with individuals within the area. Therefore, a node-link diagram that illustrate the ego-centric network of each area, and an analysis of network structures and characteristics would be beneficial to better understand the channels for information diffusion.

Sentiment analysis would be another potential direction for analyzing reciprocal communications. One can measure the impact of events on the diffusion of information by capturing the sentiment in conversational discourse. This would help predict the relative support of individuals in political campaigns, and understand how sentiment evolves through social networks and time. Future work can also be done to predict the factors that influence voter sentiment. Also, geo-social network modeling approach introduced in this article uses the number of connections for determining the geo-graphic neighbors in the neighborhood selection process. Alternatively, one can incor-porate a measure of topological similarity such as one that considers the network structure (triads), or measures such as centrality, and clustering coefficient within the neighborhood selection process.

Acknowledgments

Disclosure statement

No potential conflict of interest was reported by the author.

ORCID

Caglar Koylu 🔵 http://orcid.org/0000-0001-6619-6366

References

Adamic, L.A., *et al*. 2014. Information evolution in social networks. *arXiv preprint arXiv:1402.6792*.

Allan, J., ed., 2002. Introduction to topic detection and tracking. *In*: *Topic detection and tracking: event-based information organization*. Vol. 12. Boston, MA: Springer US, 1–16.

Andris, C., 2015. LBSN data and the social butterfly effect (Vision Paper). *In*: *Paper presented at the Proceedings of the 8th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, Seattle, WA.

Backes, M., *et al*. 2017. walk2friends: inferring social links from mobility profiles. *arXiv preprint arXiv:1708.08221*.

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

Blei, D.M. and Lafferty, J.D., 2006. Dynamic topic models. *In*: *Paper presented at the Proceedings of the 23rd international conference on Machine learning*, Pittsburgh, PA.

Blei, D.M., Ng, A.Y., and Jordan, M.I., 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.

Boyd, D., Golder, S., and Lotan, G., 2010. Tweet, tweet, retweet: conversational aspects of retweeting on Twitter. *In*: *Paper presented at the 2010 43rd Hawaii International Conference on System Sciences (HICSS)*, Honolulu, HI.

Bruns, A. and Moe, H., 2013. Structural layers of communication on Twitter. *In*: Weller, K., *et al*., eds. *Twitter and society*. Vol. 89. New York: Peter Lang, 15–28.

Bruns, A., Weller, K., and Harrington, S., 2013. Twitter and sports: football fandom in emerging and established markets. *Twitter and Society*, 89, 263–280.

Chae, J., *et al*., 2012. Spatiotemporal social media analytics for abnormal event detection and examination using seasonal-trend decomposition. *In*: *Paper presented at the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Seattle, WA.

Chakrabarti, D., Kumar, R., and Tomkins, A., 2006. Evolutionary clustering. *In*: *Paper presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia, PA.

Chen, L. and Roy, A., 2009. Event detection from Flickr data through wavelet-based spatial analysis. *In*: *Paper presented at the Proceedings of the 18th ACM conference on Information and knowledge management*, Hong Kong, China.

Chen, S., *et al*., 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. doi:10.1109/TVCG.2015.2467619

Cheng, T. and Wicks, T., 2014. Event detection using Twitter: a spatio-temporal approach. *PloS One*, 9 (6), e97807. doi:10.1371/journal.pone.0097807

Chi, Y., *et al*., 2007. Evolutionary spectral clustering by incorporating temporal smoothness. *In*: *Paper presented at the Proceedings of the 13th ACM SIGKDD International Conference on Knowledge discovery and data mining*, San Jose, CA.

Clauset, A., Newman, M.E.J., and Moore, C., 2004. Finding community structure in very large networks. *Physical Review E*, 70 (6). doi:10.1103/PhysRevE.70.066111

Cogan, P., *et al.*, 2012. Reconstruction and analysis of Twitter conversation graphs. *In*: *Paper presented at the Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, Beijing, China.

Compton, R., Jurgens, D., and Allen, D., 2014. Geotagging one hundred million Twitter accounts with total variation minimization. *In*: *Paper presented at the 2014 IEEE International Conference on Big Data (Big Data)*, Washington, DC.

Cui, W., *et al.*, 2011. Textflow: towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics*, 17 (12), 2412–2421. doi:10.1109/TVCG.2011.239

Eagle, N., Pentland, A.S., and Lazer, D., 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106 (36), 15274–15278. doi:10.1073/pnas.0900282106

Gad, S., *et al.*, 2015. ThemeDelta: dynamic segmentations over temporal topic models. *Ieee Transactions on Visualization and Computer Graphics*, 21 (5), 672–685. doi:10.1109/tvcg.2014.2388208

Gamio, L. and Borchers, C., 2016. A visual story of Donald Trump dominating the news cycle. *The Washington Post* [online], 21 July. Avaliable from: https://www.washingtonpost.com/graphics/politics/donald-trump-vs-hillary-clinton-in-media/

Gerber, M.S., 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61, 115–125. doi:10.1016/j.dss.2014.02.003

Goldstone, A. and Underwood, T., 2012. What can topic models of PMLA teach us about the history of literary scholarship. *Journal of Digital Humanities*, 2 (1), 39–48.

Grant, C.E., *et al.*, 2011. Online topic modeling for real-time Twitter search. *In*: *Paper presented at the TREC*, Gaithersburg, MD.

Guo, D. and Zhu, X., 2014. Origin-destination flow data smoothing and mapping. *Visualization and Computer Graphics, IEEE Transactions On*, PP (99), 1–11. doi:10.1109/TVCG.2014.2346271

Han, S.Y., Tsou, M.-H., and Clarke, K.C., 2018. 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*, 11 (5), 451–469. doi:10.1080/17538947.2017.1330366

Hong, L., Convertino, G., and Chi, E.H., 2011. Language matters in Twitter: a large scale study. *In*: *Paper presented at the ICWSM*, Barcelona, Spain.

Hoque, E. and Carenini, G., 2016. Interactive topic modeling for exploring asynchronous online conversations: design and evaluation of ConVisIT. *Acm Transactions on Interactive Intelligent Systems*, 6 (1), 1–24. doi:10.1145/2854158

Hu, B. and Ester, M., 2013. Spatial topic modeling in online social media for location recommendation. *In*: *Paper presented at the Proceedings of the 7th ACM conference on Recommender systems*, Hong Kong, China.

Hu, J., *et al.*, 2015. Modeling the evolution of development topics using dynamic topic models. *In*: *Paper presented at the 2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, Montreal, QC.

Huang, A., 2008. Similarity measures for text document clustering. *In*: *Paper presented at the Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)*, Christchurch, New Zealand.

Huang, Y., *et al.*, 2016. Understanding US regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, 59, 244–255. doi:10.1016/j.compenvurbsys.2015.12.003

Java, A., *et al.*, 2007. Why we Twitter: understanding microblogging usage and communities. *In*: *Paper presented at the Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, San Jose, CA.

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., *et al.*, 2012. Network analysis of three Twitter functions: favorite, follow and mention. *In*: *Paper presented at the Pacific Rim Knowledge Acquisition Workshop*, Kuching, Malaysia.

Kim, D. and Oh, A., 2011. Topic chains for understanding a news corpus. *In*: *Paper presented at the International Conference on Intelligent Text Processing and Computational Linguistics*, Tokyo, Japan.

Kim, K.S., Kojima, I., and 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

Koylu, C., *et al*., 2014. Mapping family connectedness across space and time. *Cartography and Geographic Information Science*, 41 (1), 14–26. doi:10.1080/15230406.2013.865303

Koylu, C. and 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., 2018a. Discovering multi-scale community structures from the interpersonal communication network on Twitter. *In*: L. Perez, E.-K. Kim, and R. Sengupta, eds. *Agent-based models and complexity science in the age of geospatial big data: selected papers from a workshop on agent-based models and complexity science (GIScience 2016)*. Cham: Springer International Publishing, 87–102.

Koylu, C., 2018b. Uncovering geo-social semantics from the Twitter Mention Network: an integrated approach using spatial network smoothing and topic modeling. *In*: S.-L. Shaw and D. Sui, eds. *Human dynamics research in smart and connected communities*. Cham: Springer International Publishing, 163–179.

Lai, J., Cheng, T., and Lansley, G., 2017. Improved targeted outdoor advertising based on geo-tagged social media data. *Annals of GIS*, 23 (4), 237–250. doi:10.1080/19475683.2017.1382571

Lansley, G. and Longley, P.A., 2016. The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, 58, 85–96. doi:10.1016/j.compenvurbsys.2016.04.002

Lin, Y.R., Margolin, D., and Lazer, D., 2016. 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*, 67 (9), 2072–2089.

Liu, X., *et al*. 2013. A text cube approach to human, social and cultural behavior in the twitter stream. *In*: Agarwal, N., Xu, K. S., and Osgood, N., eds. *Social computing, behavioral-cultural modeling and prediction*. Washington, DC: Springer. 321–330.

Liu, Y., *et al*., 2015. Spatio-temporal topic models for check-in data. *In*: *Paper presented at the 2015 IEEE International Conference on Data Mining (ICDM)*, Atlantic City, NJ.

Maireder, A. and Ausserhofer, J., 2014. Political discourses on Twitter: networking topics, objects and people. *In*: *Twitter and Society*. Vol. 89. New York: Peter Lang, 305–318.

Maireder, A. and Schwarzenegger, C., 2012. A movement of connected individuals: social media in the Austrian student protests 2009. *Information, Communication & Society*, 15 (2), 171–195. doi:10.1080/1369118X.2011.589908

Malik, S., *et al*., 2013. TopicFlow: visualizing topic alignment of Twitter data over time. *In*: *Paper presented at the Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Niagara, Ontario.

McCallum, A., Wang, X., and 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* [online]. Available from: http://mallet.cs.umass.edu [Accessed 30 Mar 2018].

McGee, J., Caverlee, J.A., and Cheng, Z., 2011. A geographic study of tie strength in social media. *In*: *Paper presented at the Proceedings of the 20th ACM international conference on Information and knowledge management*, Glasgow, UK.

Mok, D., Wellman, B., and Carrasco, J., 2010. Does distance matter in the age of the Internet? *Urban Studies*, 47 (13), 2747–2783. doi:10.1177/0042098010377363

Newman, D., *et al*., 2010. Automatic evaluation of topic coherence. *In*: *Paper presented at the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2–4 June, Los Angeles, CA. Stroudsburg, PA: ACM, 100–108.

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

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

Salton, G., and McGill, M., 1983. *Introduction to modern information retrieval*. New York: McGraw-Hill.

Schofield, A. and Mimno, D., 2016. Comparing apples to apple: the effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics*, 4, 287–300.

Segerberg, A. and Bennett, W.L., 2011. Social media and the organization of collective action: using Twitter to explore the ecologies of two climate change protests. *The Communication Review*, 14 (3), 197–215. doi:10.1080/10714421.2011.597250

Smith, M.A., *et al.*, 2014. Mapping Twitter topic networks: from polarized crowds to community clusters. *Pew Research Center*, 20, 1–56.

Stefanidis, A., *et al.*, 2013. Demarcating new boundaries: mapping virtual polycentric communities through social media content. *Cartography and Geographic Information Science*, 40 (2), 116–129. doi:10.1080/15230406.2013.776211

Steiger, E., Resch, B., and Zipf, A., 2016. Exploration of spatiotemporal and semantic clusters of Twitter data using unsupervised neural networks. *International Journal of Geographical Information Science*, 30 (9), 1694–1716. doi:10.1080/13658816.2015.1099658

Tamburrini, N., *et al.*, 2015. Twitter users change word usage according to conversation-partner social identity. *Social Networks*, 40, 84–89. doi:10.1016/j.socnet.2014.07.004

Tang, D., *et al.*, 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. *In*: *Paper presented at the ACL (1)*, Baltimore, MD.

Wang, X. and McCallum, A., 2006. Topics over time: a non-Markov continuous-time model of topical trends. *In*: *Paper presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, Philadelphia, PA.

Wei, W. and Carley, K.M., 2015. Measuring temporal patterns in dynamic social networks. *ACM Transactions on Knowledge Discovery from Data*, 10 (1), 1–27. doi:10.1145/2749465

Xu, T., *et al.*, 2008. Evolutionary clustering by hierarchical dirichlet process with Hidden Markov state. *In*: *Paper presented at the Eighth IEEE International Conference on Data Mining, 2008. ICDM'08*, Pisa, Italy.

Yamaguchi, Y., Amagasa, T., and Kitagawa, H., 2013. Landmark-based user location inference in social media. *In*: *Paper presented at the Proceedings of the first ACM conference on Online social networks*, Boston, MA.

Yan, X., *et al.*, 2013. A biterm topic model for short texts. *In*: *Paper presented at the Proceedings of the 22nd international conference on World Wide Web*, Rio de Janeiro, Brazil.

Zhang, D., Zhai, C., and Han, J., 2009. Topic cube: topic modeling for OLAP on multidimensional text databases. *In*: *Proceedings of the 2009 SIAM International Conference on Data Mining*. Sparks, NV: Society for Industrial and Applied Mathematics, 1124–1135.

Zhang, J., Ghahramani, Z., and Yang, Y., 2005. A probabilistic model for online document clustering with application to novelty detection. *In*: *Paper presented at the Advances in Neural Information Processing Systems*, Vancouver, BC.

Appendix

Table A1. Election-related mention topics and sample tweets in 10/15/2016–11/15/2016.

Mention tweets	T10	T35	T49
T10: Trump, white, racist, country, president, vote, America, women, black, agree, Hillary, support, voted, point, understand, ppl, change, opinion, wrong, respect			
T35: Trump, Hillary, Clinton, Obama, emails, HRC, media, president, Bush, election, corrupt, Donald, FBI, GOP, campaign, lies, rape, news, lying, Russia			
T49: Vote, Trump, election, voted, voting, Hillary, won, votes, win, party, president, Bernie, electoral, Clinton, college, polls, candidate, voters, popular, elections			

	Topic Probability		
Mention tweets	**T10**	**T35**	**T49**
A: @B if you're in our country illegally; what he says shouldn't matter. They not suppose to be here in the first place A: he's talking about only the illegal immigrants B: @A but you see how that can be confused with 'all Mexicans are...' ?? It's wrong. Trump used to fear tactics to swindle you A: @B i get what you're saying; I just don't think he was referring it to anybody besides the illegal immigrants; idec who wins lol B: @A ehh; even if he wasn't. His word choice was very divisive. But _____ ___; I do. If Trump win; we in trouble !!!	0.94		
A: but @B you can't make that generalizations. Racism works on all sides; and I've met dems who are just as prejudice. B: @A everyone has their own prejudices but that doesn't make us racist; by definition. B: @A I have never met a racist Democrat nor have I ever met a Trump supporter who wasn't a racist. A: @B so then that plays on both sides. By no means am I saying racism doesn't exist; but I'm saying that while there are prejudices A: @B and racists on both sides and both can be found in both parties. A: @B I challenge that. While I don't agree with Trump or his platform or actions but you can't generalize and say any one who A: @B supports him is racist. B: @A if someone can vote for him after the things he has said about various groups of people; then they're racist and full of hate.	0.92		
A: @B you don't have to. There are Muslims who are considering going against their religion to avoid hate crimes against them B: @A no decent person just changed their mind about sexual assault or about discrimination over night B: @A The people who think that's okay; would still have that mind set if Clinton was elected. Sorry but I don't agree. A: @B There are Muslims who have already done that!! Since 9/11!! Trump is not to blame for people being racist _____ A: @B did Trump say some racist things about Muslims? YES. But he isn't the core problem. And Hillary is not a solution. B: @A once again I disagree.	0.87		
A: @B Truth doesn't care if your female or male! She lied and that's the truth! B: @A So trump did not lie? please explain that logic; you whites has handed over a regime and tht is now terrorizing jews. B: White Americans like you @A will be responsible for destruction on this country and people lives. B: @A Please show where trump did not lie	0.84		
A: @B I'm sorry I'm a failure B: @A i wa being serious A: you must not know what a state of emergency is then lmao @B A: at least I voted @B B: so did I? @A A: @B i was mainly referring to the 11;000 plus people who voted for a dead gorilla. I was just saying I love Obama; please dont go B: I love Obama too @A			0.55

(*Continued*)

Table A1. (Continued).

Mention tweets	T10	T35	T49
A: @B I'm racist??? Lol you must not know me A: @B ahhhh not just black people are saying fuck trump lmfao get it together girl B: @A yes yes and yes. Your statement was racist PERIOD before you try to insult the intelligence of blacks get your facts straight! B: @A I have more than 1 black friend; lol be direct girl. A: @B whoaaaaa don't wanna be black but good try. I'm proud of my nationality. I'm Puerto Rican babygirl	0.5		
A: you voter Id is on here_____ @B A: just looking out for you. @B B: @A didn't you vote for Bernie sanders? A: @B Nope; voted Hilary because I prefer her over stupid Trump. Voting for Bernie would have divided the votes against trump. Sadness			0.84
A: @B Not necessarily what I'm saying. Just who casted those different votes. Plenty of Obama-Trump voters. B: @A Trump won critical battleground precincts that Obama won two elections in a row. Hillary was just a bad candidate.		0.92	
A: @B @C she got destroyed 2nd and 3rd debate. Lmao I could care less what liberal media polls say B: @A i dont think you understand how scientific polls work @C A: @B the same scientific polls that said brexit wouldn't happen and that Trump wouldn't win the nomination? Lol good try A: @B brexit is nothing like this election		0.85	
A: @B Hillary won the popular vote A: then why did you say that the majority of the country wanted trump @B B: @A Worded badly. While she won the popular vote; he won a majority of states and swing states.		0.84	0.06
A: @B As a Sanders supporter who was told 2 get over the DNC rigging the primary; Hillary supporters r being whiny & hypocritical B: @A I wish Hillary was as good at rigging things as was theorized…we could use a Corrupt Bargain about now. A: @A The DNC emails prove it; as do the ones showing Brazile gave Hillary debate quesstions: A: @B Sanders polled better against Trump too; but they had to have their pro-fracking; pro-pipelines; corrupt warmonger. B: @A Look; I don't want to argue anymore. You did it. You won. Kick us wonks out and fight populism with populism. Best. A: I wasn't arguing; just pointing out what Hillary supporters had done and what they are doing. Peace. A: @B Peace; dude. We're still on the same side..:) B: @A Peace to you; too. B: @A Here here! To 2018! I think the Sanders org will be a midterm game-changer. A: @A You know it man…:)	0.15	0.27	0.44
A: @B You are wrong; NEVER; he absolutely never said he thought the election was rigged; even if it was rigged in Florida. A: @B It is how you frame a question. Gore's challenge wasn't about 'rigged elections' it was about counting votes in Miami Dade. A: @B It is not semantics; I don't think you truly understand what 'not accepting' the election results as Donald Trump just said. A: @B It also matters when it is said BEFORE the election or AFTER the election. Donald Trump is preemptively saying it is rigged. B: @A Dude…I asked a ? Put your undies back in place. But didn't Gore challenge the results in court? That's not accepting to me. B: @A So; semantics. One says it and people get upset. The other 'challenged' the outcome; and that's okay? B: @A How many other ways is there to say what he said? Both sound like losers. B: @A Well…there are the DNC emails your candidate was involved in. Look like those things really happen. B: @A It clearly pointed to Hillary being their choice to face the GOP; no matter how much was better B: @A Cheating is Cheating. Fix is Fix. Kinda Black and White to me.	0.38	0.44	0.08