Analyzing Social Media Relationships in Context with Discussion Graphs

Emre Kıcıman, Munmun De Choudhury,
Scott Counts, Michael Gamon
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
{emrek, munmund, counts,
mgamon}@microsoft.com

Bo Thiesson*
Aalborg University
Department of Computer Science
Selma Lagerlöfs Vej 300
DK-9220 Aalborg East, Denmark
thiesson@cs.aau.dk

ABSTRACT

We present discussion graphs, a hyper-graph-based representation of social media discussions that captures both the structural features of the relationships among entities as well as the context of the discussions from which they were derived. Building on previous analyses of social media networks that focus on the strength of relationships between entities, our discussion graphs explicitly include contextual features such as who is participating in the discussions, when and where the discussions are occurring, and what else is being discussed in conjunction. There are two contributions of this work. First, we extend standard hyper-graph representations of networks to include the distribution of contexts surrounding discussions in social media networks. Second, we demonstrate how this context is useful for understanding the results of common graph measures and analyses, such as network centrality and pseudo-cliques, when applied to the analysis of textual social media content. We apply our framework across several domains captured in Twitter, including the mining of peoples' statements about their locations and activities and discussions of the U.S. 2012 elections.

1. INTRODUCTION

Much prior research has focused on analyzing social media in a large number of disparate domains, from public health such as disease modeling [20] and disease propagation [22] to crisis and violence situations such as the Mexican drug war [17], and for urban informatics such as inferring the boundaries of neighborhoods [3]. While the analytical techniques in these prior works may be varied, at their core each focused on the relative strength of pair-wise relationships mined from social media. For example, Paul and Dredze studied the relationship between textual representation of ailments and geography. Monroy-Hernández et al. examined the relationship among hashtags and user behaviors in tweets about violence related to the drug war in Mexico. Sadilek et al. modeled disease contagion (flu) by inferring relationships between people based on the locations they visited.

Critical to interpreting these relationships is the context of the social media discussions from which they are extracted. Such context can include not only aspects of who is interacting with whom, but also a variety of different attributes: the temporal (when), the spatial (where), the topical (what) dimensions, as well as the kind of mood being shared. Such context can provide insights into the underlying phenomena and suggest further lines of investigation and action. For example, in [22], understanding the kinds of locations that are more likely to spread disease from sick to healthy people may suggest possible remediations.

In this paper, we show that such analyses of social media can be re-framed as hyper-graph analyses with context information encoded on the edges. A hyper-graph representation allows us to apply common graph analyses, including network centrality and neighborhood detection, which can provide critical insights into in-domain analyses. This mapping of social media problems to their graph representations ensures that the context of the original discussions is preserved.

These hyper-graphs, which we call discussion graphs, form an analysis framework that flexibly represents both the strength and context of relationships mined from social media discussions. Discussion graphs capture the host of multi-dimensional relationships among extractable features in social media discussions, including content features, such as entity mentions, hashtags, embedded links and sentiment, message features such as when and where a message was written, and author features such as gender, popularity and expertise. Each node in the hyper-graph represents one of these typed feature values, and hyper-edges represent their co-occurrences within individual social media messages. In addition, each hyper-edge can be associated with a set of statistics that provide a summary representation of the contexts in which a particular relationship has been observed. Figure 1 shows a sample sub-selection of a discussion graph of activities and locations.

We have used discussion hyper-graphs as a framework for analyzing social media data across two distinct domains: (1) location-activity mining, and (2) politics. Our usage of this

*Work done while author was at Microsoft Research

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Eleventh Workshop on Mining and Learning with Graphs. Chicago, Illinois, USA
Copyright 2013 ACM 978-1-4503-2322-2 ...$15.00.

framework has followed a simple procedure. We first build a high-dimensional discussion graph from a social media corpus. To extract and analyze a specific set of relationships, we project the multi-dimensional hyper-graph to a lower-dimensional space, keeping only nodes relevant to our analysis, as well as the hyper-edges connecting those nodes. The values of all other features are aggregated, and the resulting conditional distributions are attached as contextual statistics to each hyper-edge and node in the lower-dimensional graph. Then, to answer our domain-specific questions, we explore the projected discussion graph directly, and we apply higher-level graph and machine learning analyses to summarize its properties. Based on our results, we frequently iterate by returning to the high-dimensional discussion graph, and re-projecting to a different set of lower dimensions to enable analysis from a different viewpoint.

In the rest of this paper, we present a more formal definition of discussion graphs, describe our tool for extracting and manipulating discussion graphs, and present two case studies applying higher-level graph analyses.

2. BACKGROUND LITERATURE

There has been considerable work around the areas of social media context mining, as well as network and structural analyses of social networks. Since our work builds on both the development of a unified model of social actions and relationships, we review literature from both topics.

2.1 Social Media Context Mining

Leveraging social media context and co-occurrence relationships between actions and attributes has been gaining considerable traction recently. It has spanned a variety of problems—social search and ranking [24], recommender systems [13, 7], community detection [19], media summarization [15], topic modeling [16, 11], and so on.

For improving entity search and ranking in social media, Weng et al. proposed TwitterRank, combining user influence and topic context [24]. Huang et al. modeled and predicted aggregates of social actions of individuals by making use of network structure, i.e., friendship links and the complex dynamics of their behavior [9]. Cranshaw et al. analyzed the geo-temporal traces of individuals and their relationships in social media to devise a set of location-based features to characterize a geographic region [3].

Researchers have derived value by making use of social context in media summarization and visualization as well. For instance, Lin et al. presented an algorithm to discover multi-relational structures from social media streams, deriving interrelationships among time, visual content, users, and actions on Flickr [15]. Social context has also been observed to be a useful signal in topic modeling. Kataria et al. modeled the influence of cited authors along with the interests of citing authors in citation networks [11].

While all of these works provide us insights into the utility of mining social media context around a variety of applications and tasks, understanding their impact on the network structure and dynamics in the light of co-occurrences of contextual features remains challenging. For instance, while cohorts of democrats and republicans on Twitter may have different topical interests and posting behavior, understanding of their social relationships is incomplete without an understanding of the graph structure that is induced as a consequence of these differences in their behavioral context. Addressing this challenge forms a core contribution of our research work.

2.2 Graph Analyses of Social Networks

Analyzing the structure and dynamics of networks within social media has also been of interest to research over the past several years.

In order to study the temporal evolution of interaction graphs, Asur et al. presented an event based characterization of behavioral patterns and the flow of information among actors over time [1]. Bakshy et al. examined the role played by social network structure (Facebook) in the diffusion of information [2]. Examining the impact of language and linguistic style on social interactions and graph structure, Danescu-Niculescu-Mizil et al. investigated power differences among Wikipedia contributors and members of the US Supreme Court [4]. Close to our work is the work of Lin et al. wherein the authors proposed a hyper-graph factorization method to detect community structure in rich media social networks, and observed how it evolves over time, through analysis of multi-relational data—topics, time, and interactions [14].

The value of making sense of the network structure on social media is tremendous, as is demonstrated by this line of research. However structure and dynamics of social graphs are often contextual. For instance, certain network structures might be conducive to only certain topics or time periods. Similarly, communities can be topical or geographical—a mountain biking community in Seattle may demonstrate different behavior than a hiking community in New York, despite having similar network representation. Marrying context with relationships addresses such nuances of our behavior, and constitutes a novel contribution of our current work.

3. DISCUSSION GRAPH

A *discussion graph* is a hyper-graph representation of a set of relationships and their associated contexts, extracted from a social media corpus [1]. A *hyper-graph* is similar to a graph, except that hyper-edges mutually connect any number of nodes, whereas edges in a graph each connect exactly two nodes. In a discussion graph, each hyper-edge is annotated with a statistical representation of the original context from which the hyper-edge was inferred.

For example, if we have many tweets that mention both the activity "hiking" and the location "tiger mountain," we can build a discussion graph where an edge connects the "hiking" node with the "tiger mountain" node. This edge is then annotated with the contextual statistics of the original tweets, such as the gender distribution of tweet authors, the time-of-day the messages were posted, and even the positive or negative sentiment expressed in the tweets. The explicit representation of context allows us to go deeper in analyzing and interpreting the relationship between "hiking" and "Tiger Mountain."

A key advantage of using a hyper-edge representation for the relationships in our discussion graph is that we can represent and analyze complex relationships. For example, we can decide to analyze the relationship between "hiking" and "Tiger Mountain" conditioned on the gender of the tweet

[1]We use the term *discussion graph* instead of *discussion hyper-graph* for brevity.

author. in this case, we create a discussion graph that is *projected* onto the 3 domains, location, activity, and also gender. Now, this discussion graph will include two hyper-edges, one which connects "hiking," "Tiger Mountain," and the "male" gender node, and another which connects the activity and location with the "female gender" node. The context associated with the former hyper-edge shows us a statistical representation of original discussion by men on this topic, and the context of the latter shows us the statistical representation of the original discussion by women. We can now quickly compare and contrast to find the gender differences in sentiment, time, word distributions, etc., that surround hiking at Tiger Mountain.

In the rest of this section, we present a formal definition of discussion graphs, then describe in detail the steps of our analysis procedure: building an initial discussion graph from social media; and focusing on specific relationships by projecting the hyper-graph representation to lower dimensions.

3.1 Formal Definitions and Notation

Formally, a hyper-graph $\mathcal{H} = (N, E)$, where N is a set of nodes and E is a set of distinct hyper-edges such that for all $e \in E$, $e \subseteq N$. A *discussion hyper-graph* extends the notion of a hyper-graph by explicitly defining statistics for each hyper-edge in the graph. Hence, a discussion hyper-graph

$$\mathcal{G} = (N, E, S)$$

, where S are the statistics associated with the edges in E — one specific $s \in S$ for each $e \in E$.

The domain D from which the discussion hyper-graph is constructed becomes important when defining the basic operations on the graph and can be thought of as the stochastic variables for which values are encoded in the graph. We will make this domain dependence explicit, by representing a discussion hyper-graph as

$$\mathcal{G}^D = (N^D, E^D, S^D),$$

Note that $S^D = \emptyset$ for the initial discussion hyper-graph

In fact, S^D will only involve statistics *not* associated with any of the variables in the domain D.

3.2 Building Discussion Graphs from Social Media

Let a social media corpus $\mathcal{C}$ be composed of a set of messages, $\mathcal{M} \equiv m_1, m_2, \ldots$. Each message m includes one or more textual components, as well as metadata about the author, embedded links, etc. Each message m_i is also identifiable by a unique identifier i.

A message is parsed by a set of low-level feature generator functions $F_d(m) = \{f_d(m)\}_{d \in D}$, where each function $f_d(m)$ may (or may not) extract, derive, or uncover a value for some feature domain d. A feature-node in the discussion hyper-graph is associated with each feature function that successfully produces a value and this node will be uniquely identified by its domain and value. Depending on the semantics of the relationship between a feature-node and a message, we may sometimes say that a node was *mentioned in*, *derived from*, or *related to* the message, the message's author or the message's context. Note that the same feature-node may be related to multiple messages.

Each message in a social media corpus creates such a hyper-edge, and the resulting multi-dimensional hyper-graph

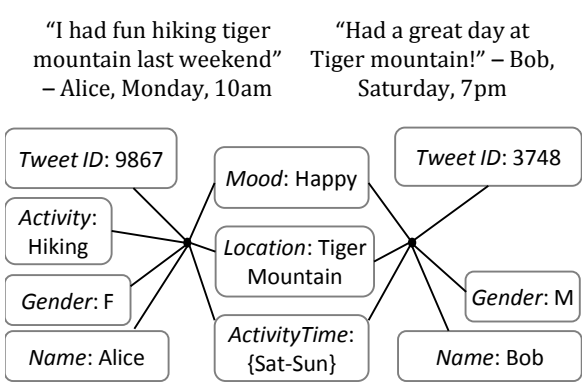

Figure 1: Example of a simple discussion showing relationships between sentiment, location, activity, post, and a variety of user attributes, such as name and gender.

is a loss-less representation of all the features extracted from the corpus. For example, Figure 1 shows a sample discussion graph extracted from two tweets.

Recall that the initial discussion hyper-graph is generated from the low-level feature functions F_D on a single message $m \in \mathcal{M}$. Here, N_m denotes a set of nodes (feature values) extracted from the message m. Hence, N_m is the union of output values produced by the functions $F_D(m) = \{f_d(m)\}_{d \in D}$. That is, $f_d(m) = n_m \in N_m$ iff $f_d(m) \neq null$. All nodes produced by the same message are interrelated on an equal footing and in that way defines the hyper-edge e_m between all nodes in N_m. Hence, $E_m = \{e_m\}$. The initial hyper-edge will not have any associated statistics, leaving $S_m = \emptyset$. We will see in the following section that the mid-level projections will add statistics to hyper-edges in the graph.

The initial discussion hyper-graph generated from the entire corpus $\mathcal{C}$ is now defined as

$$\mathcal{G}_\mathcal{C} = \bigcup m \in \mathcal{M} \mathcal{G}_m = (N_\mathcal{C}, E_\mathcal{C}, S_\mathcal{C}), \quad (1)$$

where $N_\mathcal{C} = \bigcup m \in \mathcal{M} N_m$, $E_\mathcal{C} = \bigcup m \in \mathcal{M} E_m$, and $S_\mathcal{C} = \bigcup m \in \mathcal{M} S_m = \emptyset$. In the rest of this section, we will assume we are operating on a fixed corpus and will therefore drop the subscript $\mathcal{C}$ to simplify the notation.

3.3 Hyper-graph Projection and Aggregation

In the context of a specific analysis or application, we often want to limit our structural analysis to the relationships among nodes in a small number of domains. Informally, projecting a discussion graph down to those domains consists of restricting the structure of the original graph to the given domains, and aggregating all other domains in the original discussion as contextual statistics to be associated with the edges in the new, projected discussion graph.

More formally, a *projection* $\mathcal{G}^{D \downarrow D'}$ from $\mathcal{G}^D$ down to $\mathcal{G}^{D'}$, $D' \subseteq D$ is defined in two steps. First, a temporary (improper) hyper-graph

$$G^{D \Downarrow D'} = (N^{D \Downarrow D'}, E^{D \Downarrow D'}, S^{D \Downarrow D'})$$

is constructed by removing all nodes with domain $D \backslash D'$ from the hyper-edges in E^D. Notice that a restriction oper-

ation may produce duplicate hyper-edges and therefore an improper hyper-graph. For each restricted edge, $e^{D \Downarrow D'} \in E^{D \Downarrow D'}$, we augment the corresponding statistics as

$$s^{D \Downarrow D'} = t \cup s^{D}$$

where t is the initial statistic for all the nodes we removed:

$$t = l(e^{D \setminus D}, s^{D})$$

In the second step, the projection is finalized as the hyper-graph

$$\mathcal{G}^{D \downarrow D'} = (N^{D \downarrow D'}, E^{D \downarrow D'}, S^{D \downarrow D'}),$$

constructed by reducing the graph to include only unique edges, such that

$$E^{D \downarrow D'} = \{e^{D \downarrow D'} = e | e \neq ffore, f \in E^{D \Downarrow D'}\},$$

While reducing the graph to its unique edges, we also aggregate the associated statistics of the reduced edges, using a commutative and associative aggregator function.

Note that it is often the case that the initializer used in the first step of the projection is ignored (*i.e.*, produces the statistic $t = null$). In this case the new statistics are therefore just the continued aggregation of statistics from previous projections.

3.4 Pseudo-code for Analysis Procedure

While the more mathematical notation above is useful for reasoning about the properties of discussion graphs and the operations that manipulate them, we use a succinct and easy-to-read pseudo-code representation to express the operations of a given analysis. Our analysis procedure is specified in 3 pieces:

- One `DATASOURCE` assignment specifies what social media corpus is being analyzed. Important selection features, such as date ranges, are declared here.
- One `EXTRACT` statement lists the set of feature extractors that will be applied to the social media corpus. Feature extractors optionally accept arguments that control their behavior. Some feature extractors are labeled as `PRIMARY` features. A primary feature extractor essentially acts as an input filter. Only social media messages that generate at least one primary feature will be included in the result.
- One or more `PROJECT TO` statements create projections of the discussion graph that focus on relationships important for a given analysis or application goal.

Examples of our pseudo-code analysis specification are shown below in Figures 2 and 5.

4. IMPLEMENTATION

In this section, we briefly describe our implementation of discussion graphs, including the feature extractors we apply to social data.

4.1 Framework Implementation

We implement the core of our framework as a compiler that compiles the analysis specification for a discussion graph processing job, similar to the pseudo-code presented above, into a distributed data processing program built to run on a high-level Map-Reduce-like system similar to Pig, SawzAll or Dryad [18, 21, 10]. The analysis description specifies a social media corpus to be analyzed, declares a set of features to be extracted from the messages in the corpus, and one or more projections to apply. The compiler generates an optimized implementation of the analysis, including caching and reuse of intermediate files from previous analyses. The compiled program can be tested on a single machine or run at full-scale on a cluster.

The framework provides flexibility to build discussion graphs from a variety of message-oriented social media data sources. New data sources can be analyzed by providing a plug-in for loading the data format and a mapping from the new data sources schema to a canonical schema provided by our system. New data sources do not need to have a 1-to-1 mapping to the canonical schema. However, mismatches in the schema may limit which feature extractors can be applied. For example, our gender extractor, described below, relies on the availability of the message author's name.

In addition to the functionality of building and manipulating discussion graphs, the system provides basic support for debugging and validation of extracted features and derived relationships and structures. For example, all of the edges and nodes of projected discussion graphs are annotated with a random sample of the original raw social media messages from which they were derived. This allows a user to quickly get an end-to-end view, from the raw messages to the aggregated statistics, to better understand and interpret the aggregated statistics and structures of the discussion graph. In particular, we have found this invaluable for validating the results of feature extractors and interpreting ambiguities (e.g., of hashtags or other words and sentiment), and providing insights into new features that should be extracted from messages.

In addition to our compiler, we have also implemented a web-based viewer to allow easy exploration of a discussion graph. In addition, we have built a suite of utilities to apply high-level analyses to a discussion graph, while tracking the context associated with the resulting aggregate structures. Later in this paper, we will discuss two of these high-level analyses.

4.2 Feature Extractors and Aggregators

Our framework supports an extensible set of feature extractors and aggregators. Each feature extractor is responsible for analyzing a single message at a time and, for each message, outputs zero or more detected, inferred or extracted features in one or more domains. Feature extractors are written in .NET to a common API. Extractors can be built to analyze fields in the systems canonical schema as well as extended fields in the schema of specific data sources. The extractors that we have built include:

Author statistics extractor: For each social message, this extractor generates features that represent the authors name, user id, follower count, followed count, and other basic information about the author.

Exact phrase extractor: Given a list of phrases, the exact phrase extractor generates a feature whenever one of the listed phrases is seen in the text of a social message. The domain of the feature is user-specified. For example, an analysis of political discussions may use one phrase extractor to search for the names of politicians, generating features

in the politician domain, and a second one to search for specified political topics, such as taxes, foreign policy-related phrases or jobs.

Gender extractor: This extractor infers the likely gender of the author of a message based on the user's first name as given in their profile. The inference is based on a data set of gender distributions given names, derived from census data and a Facebook data sample.

Hashtag extractor: This feature extractor generates a feature for every #hashtag token in the text of a social message.

Metropolitan area extractor: This extractor uses the methodology described in [12] to map an author's self-reported profile location to a likely geographic location, at the granularity of metropolitan areas. We interpret this feature as the likely hometown of the author.

Mood extractor: The mood extractor infers the mood or affect of the author given the text of the social message. This feature extractor is trained following the methodology published in [6].

Token extractor: This extractor generates a token for every word (white-space separated token) in the tweet. The aggregation of this feature essentially builds an unsmoothed unigram language model for the set of tweets.

Time feature extractor: This module generates features representing the publication time of the social message in absolute time and in seasonally adjusted relative time (e.g., hour-of-day, day-of-week, week-of-year).

We are currently working to integrate work in social media-oriented entity linking and named entity recognition into our framework [8, 23]. Aggregators are data-type-specific, though not feature-specific, plug-ins to our framework. We have two aggregators currently built. One is a discrete aggregator, appropriate for discrete-valued features, such as gender, hashtag and phrase features. During aggregation, the discrete aggregator provides a simple count of the number of instances of each discrete feature. Our second aggregator generates a histogram for continuous-valued features. This aggregator is appropriate for features such as the time feature or follower-count feature generated by our author statistics extractor. More flexible aggregators, such as kernel-density-estimation, is future work. We are currently working to make our core system and aggregators available publicly.

5. DATA PREPARATION

We now present the results of analyses applied to two separate discussion graphs, both derived from Twitter. In order to get access to Twitter data, we make use of the Firehose stream, made available to us through a contract with Twitter. The first discussion graph represents relationships between locations and activities, and the second represents relationships among politicians and political topics. We discuss the processing and generation of these discussion graphs.

5.1 Location-Activity Discussion Graph

The first discussion graph we analyze consists of the relationships among locations and activities. Similar to the example in Figure 1, we identify locations and activities mentioned in tweets, and extract other features, including gender, time, metropolitan area and mood. We project our discussion graph to focus on the relationships among locations and activities, and use the other features as context.

```
DATASOURCE = Twitter(‘‘9/15/12-10/15/12’’);
EXTRACT
  PRIMARY PhraseExtract(match:‘‘locationlist.txt’’,
                        domain:‘‘location’’),
  PRIMARY PhraseExtract(match:‘‘activitylist.txt’’,
                        domain:‘‘activity’’),
  MoodExtract(), GenderExtract(),
  MetroAreaExtract(), TimeExtract();

PROJECT TO (‘‘location’’,‘‘activity’’)
  NAME ‘‘LocationActivity.graph’’;
PROJECT TO (‘‘location’’)
  NAME ‘‘Location.graph’’;
```

Figure 2: Pseudo-code for the Location Activity discussion graph

We identify both locations and activities using exact phrase matching. To do so, we build a database of locations by extracting all Wikipedia articles that are marked with a latitude and longitude. These articles are primarily places, including cities, landmarks, attractions, companies, famous businesses and other entities that are strongly associated with a specific geographic location. We treat the title of the article as the canonical name of the location. We filter out names that are likely to be ambiguous with common words or set phrases (e.g., locations such as SUNDAY, Aren and Can) by using a simple language-modeling approach adapted from [23]. This process results in a dataset of approximately 580k locations with relatively unambiguous names.

We build our list of activities by mining a set of search query logs for manually generated carrier phrases that identify activities with high probability. For example, carrier phrases that we use include where to go to * places for *ing and *ing equipment, where the wildcard character * is intended to identify the name of an activity. Together, our carrier phrases identify a wide variety of activities such as jogging, camping, studying and clam digging. We apply simple conjugation rules to the verbs and filtered the list for ambiguities, resulting in a set of over 16000 phrases for over 5400 distinct activities.

We applied our analysis to 1 month of English Twitter data, from September 15, 2012 to October 15th, 2012 to find all tweets that mentioned a location or activity. From these tweets, we also extract mood, gender, metropolitan area and time information. We project the resulting raw hyper-graph down to two separate discussion graphs. The first discussion graph is projected to the relationships among locations and activities, while the second includes only relationships among locations. Figure 2 shows the pseudo-code for our analysis specification.

The resulting discussion graphs include 219,638 identified location nodes and 4595 identified activities. Figures 3 and 4 shows the mood distribution and sample edges for the "vacationing" activity in our graph. Other interesting information we find in this data set includes the observation that, while the most common mood associated with activities is joviality, we identify some activities, such as *eating* and *kissing*, as being guilty pleasures based on a high degree of association with both joviality and guilt.

5.2 Election 2012 Discussion Graph

The second discussion graph we analyze consists of the relationships among politicians and political issues during

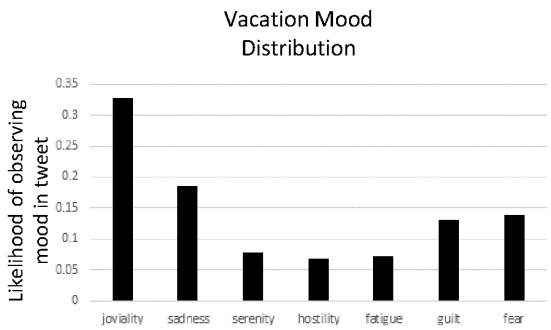

Figure 3: The mood distribution associated with the "vacationing" activity.

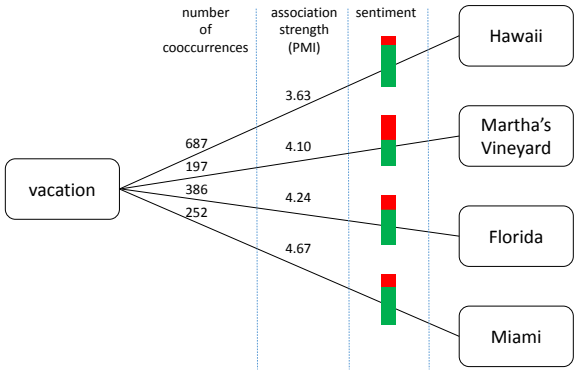

Figure 4: The strongest 4 relationships between "vacationing" and various locations, including the positive-negative sentiment context of each edge.

the last weeks of the 2012 national elections in the United States. Similar to our generation of the location and activity discussion graph, we generate our election discussion graph by using an exact phrase match to find unambiguous references to politicians and discussion topics. Figure 5 presents our analysis procedure.

6. CASE STUDY #1: CONTEXT AND PSEUDO-CLIQUES OF LOCATIONS

People discuss locations (co-mention locations) for many reasons. A person might mention two places because they are going to visit both together (e.g., "I am going to Fisherman's Wharf and then the Ferry Building"); because the two locations are comparable in some way (e.g., "The Empire State Building and Burj Khalifa are both tall buildings"); or even because two locations are dissimilar (e.g., "I want to be in sunny Hawaii, but instead am freezing in Anchorage!"). Given the variety of relationships, generating an *a priori* semantic interpretation for any given relaxed clique—essentially, understanding why a set of nodes are related to one another—is challenging if the only information we know is the existence of *some* relationship between the nodes. By looking at the contextual statistics, however, we can look for the commonalities in the nature of the relationships among the nodes in the group to characterize the nature of the set as a whole.

```
DATASOURCE = Twitter(''10/25/12-11/06/12'');
EXTRACT
  PRIMARY PhraseExtract(match:''politicianlist.txt'',
                        domain:''politician''),
  PhraseExtract(match:''issuelist.txt'',
                domain:''issue''),
  MoodExtract(), GenderExtract(),
  MetroAreaExtract(), TimeExtract();

PROJECT TO (''politician'')
  NAME ''Politician.graph'';
```

Figure 5: Pseudo-code for the Election 2012 discussion graph

In this section, we search for pseudo-cliques in a discussion graph of locations to find closely related locations. Then, we use the contextual statistics associated with the edges in the clique to characterize these pseudo-cliques and the differences between them.

6.1 Pseudo-Cliques

Intuitively, a pseudo-clique is a set of nodes that are densely connected. Together the nodes essentially form a clique with some small number of edges removed. More formally, each pseudo-clique consists of a maximal set of nodes C s.t., all nodes $n \in C$ are connected to at least $\alpha|C|$ other nodes in C. To find pseudo-cliques, we use an approximation algorithm, adapted from the algorithm presented in [5]. We calculate the context of each pseudo-clique as the aggregation of the normalized statistical distributions of the edge contexts.

6.2 Case Study #1 Results

Figure 6 shows two pseudo-cliques found in our dataset. Each of these two pseudo-clique represents a small group of locations from New York City and, in fact, the cliques share some overlap. The "Empire State Building" and "Manhattan", and "Midtown" location are members of both cliques. Given the similar locations and overlapping memberships, it is natural to question the semantic meaning of these pseudo-cliques. Is there a reason to believe these two sets of locations should be distinct?

To determine the answer, we look to the contexts associated with each pseudo-clique, shown in Figure 7, and we find that there are indeed differentiating characteristics between the two cliques: We find that the pseudo-clique on the left represents a set of relationships derived from discussions by primarily tourists, and the right represents a set of relationships derived from discussions by primarily local New Yorkers.

7. CASE STUDY #2: CONTEXT AND CENTRALITY

In this case study, we apply a network centrality measure, betweenness centrality, to a discussion graph of politicians. Recall that in our discussion graphs, politicians are related to each other if they are mentioned together in tweets. The context annotations in our graph include issues, gender, and metropolitan areas of authors. Our goal in analyzing the context associated with a discussion graph in conjunction with betweenness centrality is to gain a deeper understanding of the conditions associated with a node's centrality.

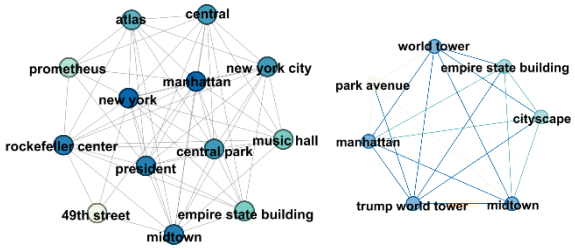

Figure 6: Two pseudo-cliques in our Location discussion graph.

		New York Tourist	Midtown Worker
Gender	Male	49%	63%
	Female	33%	23%
Metroarea	NYC	33%	54%
	Other	67%	46%
Mood	Joviality	56%	49%
	Fear	14%	13%
	Sadness	11%	15%
	Guilt	8%	6%
	Fatigue	3%	6%
	Serenity	3%	4%
	Hostility	2%	4%

Figure 7: The context of the two cliques shown in Figure 6 helps us interpret the nature of the cliques.

That is, in which contexts or situations does a node seem to play a central role, and in which does it not?

For each node n in our discussion graph, we calculate its betweenness centrality, the percentage of pairs of nodes in the graph whose shortest path passes through n. For each node, we additionally calculate two distinct contexts to associate with its centrality. First, we calculate its *positive betweenness context* as the aggregation of the contexts of distinct edges that lie on the shortest paths passing through n. In a similar fashion, we calculate the *negative betweenness context* of n as the aggregation of the contexts of the distinct edges lying on the shortest paths not passing through n. Note that some edges may contribute to both the positive and negative contexts. Given an arbitrary node m, where $m \neq n$, the more similar m's context is to the positive betweenness context of n, the more likely n lies between it and the rest of the graph. The converse holds if m's context is more similar to the negative betweenness context of n.

7.1 Case Study #2 Results

Table 1 shows the key differentiating political issues in the context associated with the two most central nodes in our discussion graph, "Barack Obama" and "Mitt Romney". Gender was not a significant differentiator, and neither were most metropolitan area values. The Pos/Neg column presents,

Obama		Romney	
Issue	$\frac{LR(Pos)}{LR(Neg)}$	**Issue**	$\frac{LR(Pos)}{LR(Neg)}$
obamacare	3.14	primary	1.71
economy	3.10	jobs	1.60
house	1.85	economy	1.56
jobs	1.60	house	1.48
senate	0.55	welfare	0.33
republican	0.28	convention	0.30
education	0.20	tea party	0.12
budget	0.19	budget	0.07

Table 1: The likelihood ratios of political issues with respect to the positive and negative betweenness contexts are shown. The table help us understand the contexts in which the Obama and Romney nodes play a central role in the discussion graph

for an issue a, the likelihood ratio of a with respect to its likelihood in the positive and negative contexts. We see that major issues in the presidential election, such as jobs and the economy (as opposed to guns, which were not a major political issue in 2012), are positively associated with both Obama's and Romney's centrality. In contrast, discussion of Republicans is negatively correlated with Obama's centrality.

8. DISCUSSION

8.1 Other Context Identifiers

In our implementation of the discussion graph framework, we chose to use a social networking message as the single, narrow object for identifying the context in which we would infer a co-occurrence relationship among two items. However, there are many other potentially meaningful objects for identifying context. For example, we might assert a relationship between two features that co-occur in the context of a longer conversation or thread of messages: if one person mentions an entity, and another person responds with a second entity, we might assert that a relationship exists. The hyper-edge relationship among features would no longer be defined by a message ID but by a conversation ID.

By selecting a different definition of the context identifier in a discussion graph, we can apply the same analytical framework to recreate a wide variety of social networking analyses. Choosing the location of a tweet as the identifying context lets us identify relationships among the people passing through the same places. Analyzing these relationships, together with context about people's health and sickness history allows us to recreate and reframe the analysis of [22] as a discussion graph analysis, simplifying the task of analyzing and comparing disease-spreading locations based on their statistical distributions of context associated with the locations and their relationships to people.

Similarly, selecting the user as a context identifier allows us to infer relationships among the locations that users visit. Applying higher-level analyses to cluster this graph of location relationships can recreate the analyses in the Livehoods project [3] while maintaining the contextual statistics embedded in a discussion graph.

8.2 Practical Experiences

In addition to the core operations of projection and ag-

gregation of discussion graphs, as presented in this paper, we have also implemented several additional operations to ease the use of discussion graphs for various applications:

Filtering: Discussion graphs can be filtered to include or exclude hyper-edges that meet certain criteria on their context or edge memberships. Depending on the analysis and application, for example, we might filter a discussion graph to include only hyper-edges that connect with at least one sentiment node, or exclude all hyper-edges with too little support. Filtering can be applied at any stage of projection, from the initial discussion graph to a final projection.

Augmentation with External Data: For some applications, it is useful to augment an inferred discussion graph with external data. For example, once we identify the name of a location, it is useful to augment the discussion-graph with a latitude-longitude feature; or to augment extracted food names with nutritional information.

Planar graph projection: For many analyses, we wish to focus on pair-wise relationships among features in a discussion graph, rather than k-way hyper-edges. A planar projection of the hyper-graph achieves this. Also frequently useful is the bipartite planar representation, where the planar graph is restricted to include only edges connecting nodes of differing domains. For example, we may not care about relationships among politicians, but only the relationships between politicians and issues.

To date, we have used our implementation of discussion graphs for a large variety of social media analyses across several domains, including analyzing the interest in and the sentiment towards nominees of movie and music awards, capturing information about people's fitness activities, the relationship among restaurants and food items, and analyzing social media usage during crises. Overall, we have found that one of the primary practical benefits of the discussion graph framework has been the ability to succinctly represent an analysis, quickly explore the results and revise our analyses to ask new questions.

9. SUMMARY

In this paper, we proposed a framework for formally modeling the context of relationships inferred from social media discussions, actions, and attributes. Through two case studies around location-activity mining and the 2012 US Presidential elections, we demonstrated how our hyper-graph representation of discussions on Twitter can lend valuable insights that are not possible with graph analyses or mining of social media actions alone. We believe that this approach to combining structural and contextual analysis of social media represents the beginning of a fundamentally new methodological direction to analyzing social media. By presenting a set of simple and flexible primitives for operating on social media data and building a system that implements these primitives, we hope to spur deeper research into social media analytics, and insights into real-world phenomena and macro-level social processes such as propagation of social influence, expertise finding, crisis mitigation or public health.

10. REFERENCES

[1] S. Asur, S. Parthasarathy, and D. Ucar. An event-based framework for characterizing the evolutionary behavior of interaction graphs. In *Proc. of SIGKDD*. ACM, 2007.

[2] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proc. of WWW*. ACM, 2012.

[3] J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proc. of ICWSM*, 2012.

[4] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg. Echoes of power: Language effects and power differences in social interaction. In *Proc. of WWW*. ACM, 2012.

[5] G. Danezis, T. Aura, S. Chen, and E. Kıcıman. How to share your favourite search results while preserving privacy and quality. In *Privacy Enhancing Technologies*. Springer, 2010.

[6] M. De Choudhury, M. Gamon, and S. Counts. Happy, nervous or surprised? classification of human affective states in social media. In *Proc. of ICWSM*, 2012.

[7] W. Feng and J. Wang. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. In *Proc. of SIGKDD*. ACM, 2012.

[8] S. Guo, M.-W. Chang, and E. Kıcıman. To link or not to link? a study on end-to-end tweet entity linking. In *Proc. of NAACL-HLT*, 2013.

[9] S. Huang, M. Chen, B. Luo, and D. Lee. Predicting aggregate social activities using continuous-time stochastic process. In *Proc. of CIKM*. ACM, 2012.

[10] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly. Dryad: distributed data-parallel programs from sequential building blocks. *ACM SIGOPS Operating Systems Review*, 41(3):59–72, 2007.

[11] S. Kataria, P. Mitra, C. Caragea, and C. L. Giles. Context sensitive topic models for author influence in document networks. In *Proc. of IJCAI*. AAAI Press, 2011.

[12] E. Kıcıman. Omg, i have to tweet that! a study of factors that influence tweet rates. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[13] I. Konstas, V. Stathopoulos, and J. M. Jose. On social networks and collaborative recommendation. In *Proc. of SIGIR*. ACM, 2009.

[14] Y.-R. Lin, J. Sun, P. Castro, R. Konuru, H. Sundaram, and A. Kelliher. Extracting community structure through relational hypergraphs. In *Proc. of WWW*. ACM, 2009.

[15] Y.-R. Lin, H. Sundaram, M. De Choudhury, and A. Kelliher. Discovering multirelational structure in social media streams. *ACM TOMCCAP*, 8(1):4, 2012.

[16] Q. Mei, X. Shen, and C. Zhai. Automatic labeling of multinomial topic models. In *Proc. of SIGKDD*. ACM, 2007.

[17] A. Monroy-Hernández, E. Kıcıman, M. De Choudhury, S. Counts, et al. The new war correspondents: the rise of civic media curation in urban warfare. In *Proc. of CSCW*. ACM, 2013.

[18] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins. Pig latin: a not-so-foreign language for data processing. In *Proc. of SIGMOD*. ACM, 2008.

[19] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos. Community detection in social media. *Data Mining and Knowledge Discovery*, 24(3), 2012.

[20] M. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *Proc. of ICWSM*, 2011.

[21] R. Pike, S. Dorward, R. Griesemer, and S. Quinlan. Interpreting the data: Parallel analysis with sawzall. *Scientific Programming*, 13(4), 2005.

[22] A. Sadilek, H. Kautz, and V. Silenzio. Modeling spread of disease from social interactions. In *Proc. of ICWSM*, 2012.

[23] C. Wang, B. Hsu, M. Chang, and E. Kiciman. Simple and knowledge-intensive generative model for named entity recognition. Technical report, Microsoft Research Technical Report, 2012.

[24] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proc. of the third ACM Intl. Conf. on Web search and data mining*. ACM, 2010.