Discussion Graphs: Putting Social Media Analysis in Context

Emre Kıcıman, Scott Counts, Michael Gamon
Microsoft Research
{emrek, counts, mgamon}@microsoft.com

Munmun De Choudhury
School of Interactive Computing, Georgia Tech
mchoudhu@cc.gatech.edu

Bo Thiesson
Dept. of Computer Science, Aalborg University
thiesson@cs.aau.dk

Abstract

Much research has focused on studying complex phenomena through their reflection in social media, from drawing neighborhood boundaries to inferring relationships between medicines and diseases. While it is generally recognized in the social sciences that such studies should be conditioned on gender, time and other confounding factors, few of the studies that attempt to extract information from social media actually condition on such factors due to the difficulty in extracting these factors from naturalistic data and the added complexity of including them in analyses. In this paper, we present a simple framework for specifying and implementing common social media analyses that makes it trivial to inspect and condition on contextual information. Our data model—discussion graphs—captures both the structural features of relationships inferred from social media as well as the context of the discussions from which they are derived, such as who is participating in the discussions, when and where the discussions are occurring, and what else is being discussed in conjunction. We implement our framework in a tool called DGT , and present case studies on its use. In particular, we show how analyses of neighborhoods and their boundaries based on geo-located social media data can have drastically varying results when conditioned on gender and time.

1 Introduction

Social media data have shown themselves to be a rich source of information about phenomena in a large variety of domains, from public health and politics to economics and urban informatics. In the health domain, for example, researchers learn potential relationships between drugs and their symptoms and side-effects based on the co-mentions of drug names and ailments in Twitter messages (Paul and Dredze 2011), and study disease transmission in the physical world by inferring transmission relationships between infected people and others based on their co-visited locations (Sadilek, Kautz, and Silenzio 2012). To study the development and growth of online communities in the context of the Mexican drug war, Monroy-Hernandez et al. study the co-occurrence relationship among hashtags and user behaviors (Monroy-Hernández et al. 2013). Cranshaw et al. extract relationships between locations based on co-visits by the same users to learn community and neighborhood boundaries in the real-world (Cranshaw et al. 2012).

Critical to a thorough exploration of relationships inferred from social media data is the *context* of the social media discussions from which relationships are extracted. Such context may include a rich set of features, such as temporal, spatial and topical context, as well as population, sentiment, and domain-specific features. Such context provides insights into the underlying phenomena and can suggest further lines of investigation and action.

Unfortunately, the effort of implementing such deep analyses around the context of social media is non-trivial. First, given the breadth of technologies involved in extracting entity, sentiment and other information from social media messages, simply finding the right models, algorithms and expertise in the first place can pose significant difficulty. Second, even after low-level features have been extracted, extracting relationships between items and tracking the context of the relationship requires a substantial programming effort (hours or days) — an effort that, to a large-degree, must be duplicated each time we extract a new kind of relationship or condition on new contextual factors.

Our goal in this paper is to drastically simplify analysis of social media data in context: to simplify the specification and implementation of common analyses, to simplify the extraction of conventionally important features such as gender as well as new features such as sentiment, and to simplify the conditioning of results on multiple kinds of context. Furthermore, we wish to make it trivial to follow best practices that improve our understanding of these extracted relationships, such as summarizing the context of the discussions from which they are extracted and tracking supporting evidence. In this paper, we focus on *co-occurrence analysis*, a large and important class of social media analyses, founded on the assumption that frequently co-occurring items may share some true relationship.

To this end, we present *discussion graphs*, a data model for representing and computing upon the relationships extracted from social media. Informally, a discussion graph is a hyper-graph representation of a set of relationships and their associated contexts. Each node in our graph represents some feature-value extracted from a social media corpus; and each hyper-edge represents a unique context (whether a specific social media message, location, time, etc.) that con-

nects all the feature-values that were found to co-occur in the same context. Each hyper-edge is annotated with a statistical summary of the discussion context from which it was derived. Simple operations on this discussion graph accommodate the wide-variety of domains to which co-occurrence analyses of social media analysis has been applied today.

To give a simple hypothetical example, we might be interested in the relationship between drugs and side-effects, as expressed in co-occurrence of drug and effect-related terms in social messages. The possible conditioning context may include gender and age of the messaging people. Our discussion graph model simplifies this kind of analysis by enabling us (1) to provide a framework for re-use of feature extractors for co-occurrence statistics and context extraction, and by (2) allowing us to project the discussion graph to the co-occurrence that we are interested in (drug $\leftrightarrow$ side effect) while aggregating statistics about the contexts (gender and age) for each relationship. To examine other conditioning contexts, all we would need is a feature extractor for that particular new context, and with a minimal change in our analysis script we could add the new context to the aggregate statistics around drug and side effect co-mentions.

We implement this data model in *DGT (Discussion Graph Tool)*, an easy-to-use analysis tool that provides a domain-specific language for operating on discussion graphs and extracting co-occurrence relationships from social media, and automates the onerous tasks of tracking the context of relationships and other best practices. DGT provides a single-machine implementation, and also generates map-reduce-like programs for distributed, scalable analyses. We have used DGT for analyses supporting production features at a large web service.

We demonstrate the applicability of discussion graphs in examples and case studies across several domains, including political discussions, relationships between activities and locations, and using context to interpret cliques of co-mentioned locations. Ultimately, we believe the primary impact of identifying a simplifying abstraction underneath co-occurrence analyses is to enable and encourage deeper analysis of social media, and especially the implications of context for interpreting and understanding relationships.

Our final case study demonstrates how conditioning on context can significantly alter the results of an analysis. We look at neighborhood boundaries learned from social media data, as in (Cranshaw et al. 2012), and find that conditioning our analysis on different contexts (day vs. night, weekday vs. weekend, male vs. female), substantially changes the inferred shape of neighborhoods. The lesson we learn is that the original context from which we learn relationships is critical to the final result of high-level analyses. By simplifying the process of extracting such contextual features and conditioning on them, DGT makes such deep analyses of dependencies feasible.

2 Background

In this section, we first discuss the goals and methodologies of social media analysis research. Then we discuss how conventional social sciences handle the conditioning of analyses on critical context and the new challenges that arise in a social media context. Finally, we briefly discuss existing social media analysis tools and how our work is complementary.

2.1 Social Media Analysis

Much social media research has focused on whether or not certain relationships exist at all in social media, as well as validating these relationships via juxtaposition with ground truth data. In the health domain, studies have looked at the relationships between diseases, medicines, side-effects, and symptoms (Paul and Dredze 2011; Myslín et al. 2013) as well as disease transmission (Sadilek, Kautz, and Silenzio 2012). Similar studies have been conducted in urban informatics (Cranshaw et al. 2012; Schwartz et al. 2013), mental health (De Choudhury, Counts, and Horvitz 2013b; Golder and Macy 2011), natural disaster monitoring (De Longueville, Smith, and Luraschi 2009; Sakaki, Okazaki, and Matsuo 2010), finance (Bollen, Mao, and Zeng 2011) and other domains. Surprisingly many of these analyses rely on a co-occurrence analysis: the assumption is that items that co-occur frequently may share some true relationship. For example, Sadilek et al.'s analysis of disease contagion infers relationships between disease carriers and new infections based on co-visited locations. Paul and Dredze studied the relationship between mentioned ailments and the geographies in which they occur. Cranshaw et al. extract the social similarity between locations based on co-visits by individuals, inferring community and neighborhood boundaries (Cranshaw et al. 2012).

Given the many biases known to be present in social media (Calais Guerra, Meira, and Cardie 2014; Kıcıman 2012), validation of information inferred from social media is critical. Many studies perform a validation by comparison with ground-truth data based on expensive conventional surveys and study techniques (Schwartz et al. 2013; De Choudhury, Counts, and Horvitz 2013b; Cranshaw et al. 2012). However, validation is especially challenging in cases where everyday behavior inferred from social media is hard to capture in traditional survey and study methods. This raises many questions around conditional factors: for whom do the effects extracted from the data apply, in what situations are the results valid, and so on. The need for easy, rapid, iterative analysis and exploration of complex contextual factors is the primary motivator for the formalization and implementation of discussion graphs presented here.

Our work also has applicability to the large body of research on improving computing and information systems. For example, social media analysis methods have been used in social search and ranking (Weng et al. 2010), recommender systems (Konstas, Stathopoulos, and Jose 2009; Feng and Wang 2012), media summarization (Lin et al. 2012), and so on. Smith et al., 2008 utilized personal social context (individuals and the communities they belong to) and community social context (individuals' information role and identity in different communities) to facilitate productive participation and search for users. Konstas et al., 2009 combined explicit multimedia-enriched data and implicit user preferences on last.fm to propose a virtual network representation for a recommender system.

2.2 Conditioning on context: person and environmental factors

Analyses in the social sciences often are conditioned on demographic variables such as gender and socioeconomic status. These variables are used as covariates to show that the phenomenon of interest explains variance in the outcome variable beyond what the demographic factors alone would explain. Including these factors when analyzing naturalistic data such as social media is challenging. As we show in our case studies, the discussion graph framework can incorporate these variables into analyses of social media in a flexible way. For instance, our second case study shows how neighborhood boundaries drawn from social media posts differ for men and women.

Social media do have advantages in that the set of variables analogous to demographics can both extend the set of person variables on which to condition, as well as provide a set of environmental variables typically absent from traditional social science analyses. For instance, variables like time of day are almost always available in social media and can be used to condition results such as the neighborhood boundary analysis shown in our third case study. Further, many variables that require inference are available based on text classification. Our first case study contrasts mood distributions over seven mood classes for tourists and locals in New York, highlighting a factor largely novel for conditioning on in the social sciences. Thus by incorporating time, mood, and other contextual factors into the hypergraph representation of the data, the discussion graph framework simplifies exploration of covariates, both traditional and novel with respect to the social sciences.

2.3 Tools for Social Media Analysis

To perform social media analyses, researchers typically depend on a broad stack of tools for text and graph analyses, machine learning and statistical analyses. Textual and metadata analyses are often used for low-level feature extraction such as entity recognition and sentiment analysis. Network analysis tools, such as Gephi (Bastian, Heymann, and Jacomy 2009), NodeXL (Hansen, Shneiderman, and Smith 2010), and SNAP (Leskovec 2013) are available for computing graph analyses on existing network data. Our work complements these classes of existing tools by focusing on extracting co-occurrence relationships of social media features and turning this into a conditioned and contextualized representation amenable for higher-level analyses.

The most similar research effort in motivation includes Heer et al. (Heer and Perer 2011)'s Orion proposal for an interactive network modeling and visualization tool, enabling rapid manipulation of large graphs using relational operators and network analytics. Similarly, although not on social data, Liu et al. (Liu, Navathe, and Stasko 2011) devised the Ploceus system that lets users analyze multivariate tabular data through network abstractions at different levels and from different perspectives. Our paper builds in this direction and proposes an easy-to-use flexible framework to ingest, aggregate and manipulate social media features and their co-occurrence analyses.

3 Discussion Graphs and DGT

In this section, we present our data model and implementation. We present a formal definition of discussion graphs and the basic operations that support co-occurrence analysis and ease the analysis of relationships in context. Based on this, we define our domain-specific scripting language and present a simple example program.

3.1 Discussion Graph Data Model

Informally, 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 [2].

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 Basic Operations

The computational pipeline for our framework has three main stages. First, we apply a series of feature extractors to a corpus of social media messages. The resultant features create the base discussion graph. Each node in the discussion graph corresponds to a distinct feature value, and each hyper-edge corresponds to a social media message, connecting all the feature values extracted from that message. Second, we apply a set of transformations to the base discussion graph. These operations set the relationship context, filter, project and augment the discussion graph to create a derived graph representation of the co-occurrence relationships among selected domains of interest.

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

[2] Note that while annotations are applied only to hyper-edges, we can calculate the distribution for a node by summing over all hyper-edges that include the node.

Third (optionally), we may apply graph analyses to the discussion graph, such as shortest-path, network centrality and neighborhood finding algorithms; or machine learning algorithms, such as clustering or classification algorithms to the nodes and edges, and their annotations, within the graph. In case study #1, we briefly show an example of how contextual statistics can be propagated to aid interpretation of the results of a graph algorithm, and in case study #2, how a discussion graph can be analyzed using spectral clustering. We take advantage of abundantly available third-party tools, such as R, to analyze our generated discussion graphs. We leave full discussion of this third step outside the scope of this paper, but suffice it to say that we have found the derived discussion graphs to be an amenable representation for both graph analyses and many machine learning algorithms.

Building the Base Discussion Graph *What features should be extracted from the social messages?*

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 metadata. Note that the same feature-node may be related to multiple messages.

Each message in a social media corpus creates a hyper-edge that connects all nodes derived from the message. In aggregate, the resulting hyper-edges constitute a multi-dimensional hyper-graph that gives a loss-less representation of all the features extracted from the corpus. Figure 1 shows a sample discussion graph extracted from two social media messages.

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

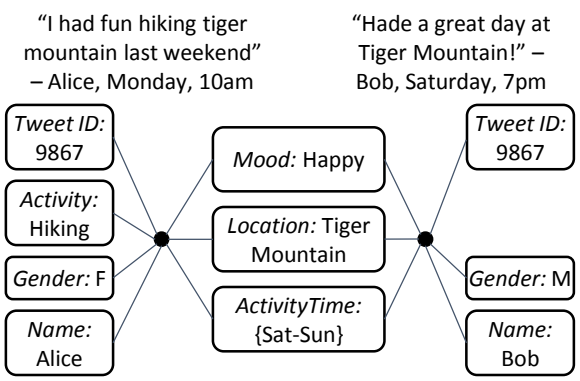

Figure 1: Example of a simple discussion showing relationships between sentiment, location, activity, post, and a variety of user attributes, such as name and gender.

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

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.

Relate By *What defines relationships?*

By default, co-occurrence in a social media message defines relationships in the initial discussion graph. That is, we will only infer co-occurrence relationships among items that co-occur in the same message. For many analyses, however, we would like to generalize the notion of co-occurrence. For example, we might want to infer relationships between locations based on the users who co-visit them. Or, we might wish to infer relationships among items mentioned across messages in a conversation. Normally, we will define this relationship before our projection.

To do so, we declare a new relationship R, and transform the hyper-graph by connecting all hyper-edges that share a common instance of a node in the domain R, such that all hyper-edges connected to a node R_i now become a single hyper-edge. Note that in some cases, a hyper-edge may now be connected to the same node multiple times. We simply treat this as a weighted connection, for purposes of aggregation in future projections.

Projection *What is the domain of relationships to extract?*

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

Command	Description
`LOAD`	Specifies the social media corpus being analyzed. Important selection features, such as date ranges, are declared here. DGT has native support for delimiter-separated text files with user-specified schemas.
`EXTRACT`	Specifies the set of feature extractors, including arguments, that will be applied to the social media corpus. A `PRIMARY` tag is a preliminary filter. Only hyper-edges including a `PRIMARY` feature are included in the resultant hyper-graph.
`RELATE BY`	Specifies the domain that defines a co-occurrence relationship.
`PROJECT TO`	Create projections of the discussion graph that focus on relationships important for a given analysis or application goal.
`OUTPUT TO`	Outputs the specified raw data or discussion graph to a specified file for analysis by R or other high-level analysis tools.

Table 1: Basic script commands

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

is constructed by removing all nodes with domain $D \backslash D'$ from the hyper-edges in E^D. Notice that a restriction operation 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 that we removed from the hyper-edge. These statistics are often just a simple function of the values represented by the removed nodes, but may, in cases, also functionally depend on the current recorded statistics in the hyper-edge.

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

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

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.

3.3 DGT and Example Usage

We implement the discussion graph data model in the DGT and illustrate its use in two examples. First, consider an analysis that involves the relationship between activities and locations — in other words *what* people do (`hiking` or `reading`) and *where* they do it (`Yosemite Falls trail` or the `Palo Alto library`). For example, if we have many tweets that mention both the activity "vacationing" and the location "Hawaii", we can build a discussion graph where an edge connects the *vacationing* node with the *Hawaii* 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 *vacationing* and *Hawaii*.

We write this example analysis using a succinct and easy-to-read script, described in Table 1.

We will go into more detail on these specific feature extractors in Section 3.4 and the activity and location scenario itself in Section 4.1. Suffice it to say that this analysis is reading from a Twitter data source, then extracting references to activities and locations, as well as several other features. From these references, the analysis is creating a graph of the relationships among all locations and activities (the `LocationActivity` graph), and also creating a graph of the relationships among locations alone (the `Location` graph). The edges in the `LocationActivity` graph are annotated with the conditional distributions of the other extracted features (the mood, gender, metropolitan area and time distributions). For example, Figure 3 shows the mood distribution associated with the *vacationing* node. The edges in the `Location` graph are augmented with those distributions as well as the activity distribution. Figure 4 shows the strongest 4 relationships, as measured by pointwise mutual information (PMI), between the *vacationing* activity and various locations. In the figure, we have annotated each of these relationships with the positive-negative sentiment associated with vacationing at that location.

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 *vacationing* and *Hawaii* conditioned on the gender of the tweet 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 *vacationing*, *Hawaii*, 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 will include 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 vacationing in Hawaii.

Once we have projected the discussion graph to a set of

```
LOAD Twitter(startdate:"9/15/12",
             enddate:"10/15/12");
var g = EXTRACT
  PRIMARY PhraseExtract(match:"locationlist.txt",
                        domain:"location"),
  PRIMARY PhraseExtract(match:"activitylist.txt",
                        domain:"activity"),
  MoodExtract(), GenderExtract(),
  MetroAreaExtract(), TimeExtract(),
MessageId
FROM s;

PROJECT TO location, activity FROM g;
OUTPUT TO "LocationActivity.graph";

PROJECT TO location FROM g;
OUTPUT TO "Location.graph";
```

Figure 2: Script for the Location-Activity discussion graph. By default, relationships are defined by co-occurrence in a social message

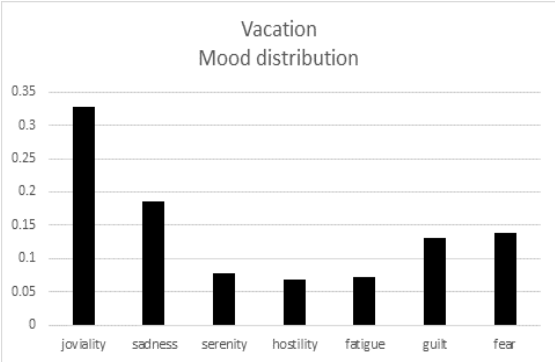

Figure 3: The mood distribution associated with the *vacationing* activity.

relations we are most interested in, we can also apply more sophisticated analyses to the results. For example, a cursory examination of the mood distributions associated with activity nodes in the `LocationActivity` graph shows us that while the most common mood associated with activities is joviality, some activities are associated strongly with both joviality and guilt, such as *eating* and *kissing*, and would be perhaps better characterized as "guilty pleasures".

As a second example, we highlight how we can project our discussion graph to "time" as a feature, in order to capture the dynamics of social media discussion. Figure 5 shows an analysis script that extracts out the time-varying political discussion associated with major politicians. The hyperedge between a politician and day will capture the distribution of issues, sentiment, gender, etc., co-occurring with mentions of the politician on each day. Figure 6 shows the timeline of issues being discussed with mentions of Obama in the weeks prior to the 2012 election.

3.4 Feature Extractors

Our framework supports an extensible set of feature extractors and aggregators, to facilitate the contribution and sharing of new extractors. 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 as DLLs, adhering to a common API. Users can extend the system with additional feature extractors to analyze fields in the system's canonical schema as well as extended fields in the schema of specific data sources. By de-

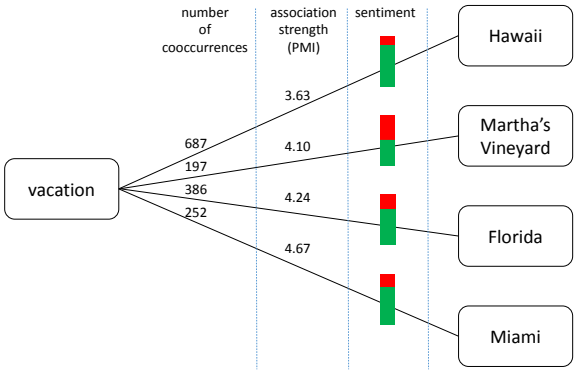

Figure 4: The strongest 4 relationships between *vacationing* and various locations, including the positive-negative sentiment context of each edge.

```
LOAD Twitter(startdate:"9/1/12",
             enddate:"11/06/12");
EXTRACT
    PRIMARY PhraseExtract(
   match:"politicianlist.txt",
   domain:"politician"),
  Time(options:``absoluteday''),
    PhraseExtract(
   match:"issuelist.txt",
       domain:"issue"),
    MoodExtract(), GenderExtract(),
    MetroAreaExtract(), TimeExtract();

PROJECT TO politician,absoluteday;
OUTPUT TO "PoliticianPerDay.graph";
```

Figure 5: Script for extracting a timeline summary of discussions about politicians.

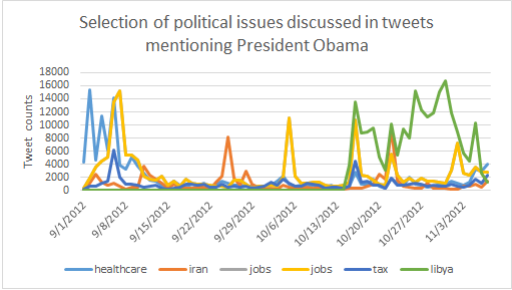

Figure 6: The distribution of selected issues co-mentioned with Obama from Sep. 1-Nov. 6, 2012. The analysis script in this case study equivalently extracts the timeline of sentiment, gender, and other summary statistics.

fault, the same feature extractor will run in both the single-machine and distributed implementation of our framework, though feature extractors may also be optimized independently for different execution platforms if necessary for achieving high performance. Some of the extractors that we have built to date include: author statistics, tokens, time features, hashtags, user mentions, phrase extractors, sentiment extractor (De Choudhury, Gamon, and Counts 2012), entity linking (Guo, Chang, and Kıcıman 2013), gender extractor (De Choudhury, Counts, and Horvitz 2013a), and hometown extractor (Kıcıman 2012).[3] Note that while we have integrated feature extractors for some common conditioning contexts, such as gender, other common factors are the topic of current research in the field (Kosinski, Stillwell, and Graepel 2013; Mislove et al. 2011).

Aggregators are data-type-specific, though not feature-specific, plug-ins to our framework. To date, we have built 1) a simple counting aggregator, appropriate discrete-valued features such as gender or hashtag; and 2) a histogram aggregator for continuous-valued features, such as time feature or follower counts.

4 Case Studies

We present two case studies with the purpose of highlighting the importance of the original contexts from which social media is captured, and that conditioning on these contexts can have substantial effects on an analysis. Exploring these results fully can be onerous if the analyses are done manually. However, with tool support from a system like DGT , filtering and pivoting the data to explore different contexts and even different kinds of relationships and relationship contexts is a matter of only a minute or two of scripting.

First, we show how contextual information captured in the discussion graph can be used to help interpret the higher-level graph structures (pseudo-cliques). Our second case study shows how higher-level analyses (inferred neighborhood boundaries) vary based on the original context in which social media was captured, with implications for how we determine which contexts or combinations are "correct".

4.1 Case Study #1: Context and Pseudo-Cliques

In this case study, we show how contextual information can be propagated to help interpret the results of a higher-level graph analysis on the relationships between co-mentioned 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 this variety, understanding why a set of nodes are related to one another is challenging—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.

To illustrate this point, we search for pseudo-cliques in a discussion graph of locations to find closely related locations. We use the contextual statistics associated with the edges in the clique to differentiate these pseudo-cliques.

Data Preparation: Location-Activity Discussion Graph The first discussion graph consists of the relationships among locations and activities. Using the script presented in Figure 2, 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 use the other features as context.

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, and hence typically represent places. We treat the article title as the canonical name of the location. We filter out names that are likely to be ambiguous with common terms using information from a large Twitter language model. The final dataset contains ˜ 580k locations.

We build our list of activities by mining search query logs for carrier phrases that reliably identify activities. Examples include patterns such as [where to go to *], [places for *ing], and [*ing equipment], where * identifies the name of an activity. This yields a wide variety of activities such as "jogging", "studying" and "clam digging". We apply conjugation rules to the verbs and filter for ambiguities, resulting in a set of over 16000 phrases for over 5400 distinct activities.

We apply our analysis to one month of English Twitter data extracting all tweets that mention a location or activity, with mood, gender, metropolitan area and time information. We project the resultant raw hyper-graph down to two separate discussion graphs: one is projected to the relationships among locations and activities, while the other includes only relationships among locations. Figure 2 shows the pseudo-code for our analysis specification. The resultant discussion graphs include 219,638 identified location nodes and 4595 identified activities.

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. We use an approximate pseudo-clique finding algorithm, and calculate the context of each pseudo-clique as the aggregation of the normalized statistical distributions of the edge contexts.

Results Figure 7 shows two pseudo-cliques found in our dataset. Each of these represents a small group of locations from NYC and 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

[3]The details of our more complex feature extractors are described in detail in the cited prior work.

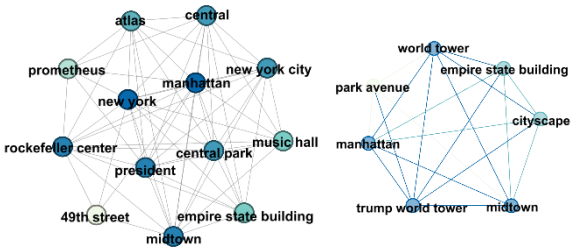

Figure 7: Two pseudo-cliques in our Location discussion graph.

		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 8: The context of the two cliques shown in Figure 7 helps us interpret the nature of the cliques. Based on this context, we manually label the cliques as "New York Tourist" and "Midtown worker"

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 8: 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.

4.2 Case Study #2: Neighborhood Boundaries in Context

In this section, we describe our case study on extracting neighborhood boundaries based on user locations observed in social media. In our previous example, we built pseudo-cliques of locations based solely on co-mentions within the text of tweets. In contrast, here we will use DGT to learn co-visit relationships among geo-locations. We follow the basic methodology outlined by the Livehoods project (Cranshaw et al. 2012) to infer a relationship between locations based on user co-visits.[4] In addition, however, we use DGT to extract these relationships conditioned on a number of different contexts, to see how much of a role various factors play. We find that neighborhood boundaries indeed change significantly from day to night, weekend to weekday and even based on gender of the twitter user.

[4]Note that, because we are using a different kind of location data, geo-located tweets instead of foursquare check-ins, as well as differences in our clustering algorithm, we are not expecting to recreate the same neighborhoods as Livehoods.

```
LOAD Twitter(startdate:"1/1/13",
             enddate:"3/31/13");
EXTRACT
  PRIMARY GeoPoint(minlatlon:"40.6 -74.1",
                   maxlatlon:"40.9 -73.8"),
  Time(options:"hourofday,dayofweek"),
userid;

// optionally FILTER to some context

RELATE BY userid;
PLANAR PROJECT TO ("geopoint");
OUTPUT TO "GeoRelationsByUser.graph";
```

Figure 9: Script for extracting relations between geo-locations based on user co-visits

Data and Analysis From our organization's archive of the Twitter firehose, we extract all geo-located tweets in the NYC area that occurred between January 1 and March 31, 2013, rounding the geo-location to 3 decimal places, resulting in 2.3M geo-located tweets. For each tweet, we extract the hour of the day, the day of the week, and the gender and userid of the author. We set our relationship context to be the userid and project our discussion graph to the relationships among geo-locations: two geo-locations are related based on the number of common userids that have visited both geo-locations. The DGT script for this extraction is shown in Figure 9. As in Livehoods (Cranshaw et al. 2012), we mix this social distance with geographic distance to build a sparse affinity matrix among locations, and cluster locations using spectral clustering.

To test our hypothesis that the neighborhood boundaries are sensitive to context, we repeat this extraction, adding one line to the analysis to filter to tweets occuring during daytime hours (7am-7pm), nighttime (7pm-7am) and weekends (Sat,Sun), weekdays. We also produce an analysis based on the context of gender. We generate neighborhood boundaries for each set of inferred social relationships independently and compare the results.

Results While there is some stability in the general pattern of neighborhoods, we also find notable differences. The neighborhood boundaries shift substantially in the midtown area of Manhattan between the day and night (Figure 10): at night, several neighborhood clusters south of Central Park merge while some new clusters emerge in the very south of Manhattan. Looking at weekend and weekday neighborhoods (Figure 11) we see that the weekend clusters reveal the 5th Ave shopping area as a distinct cluster, and show a distinction between a northern Central Park cluster and a neighboring Yorkville cluster that is not apparent during the weekdays. The gender context (Figure 12) illustrates a

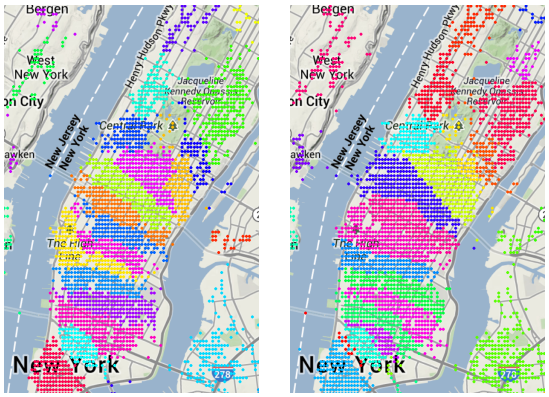

Figure 10: Neighborhoods based on day (left) and night (right) behaviors

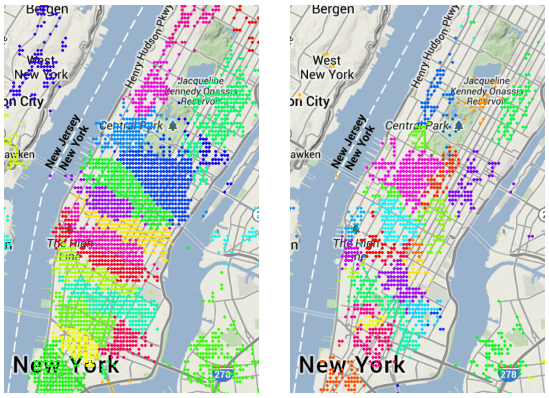

Figure 11: Neighborhoods based on weekday (left) and weekend (right) behaviors.

much stronger differentiation of neighborhoods in the midtown area, and a clear distinction between the Greenwich Village and the Soho Shopping District for the female population, whereas these distinctions are absorbed in larger clusters in the male context.

To measure the relative impact of each of these three factors on the final clustering results, we apply a pair-counting $F_{0.5}$-measure to each set of conditioned clusterings, where $F = 0$ indicates no similarity and $F = 1$ indicates identical clustering assignments.[5] We find that conditioning on weekend vs. weekday has the largest effect (0.47); gender has the second largest effect (0.54); and day vs. night has, relatively, the least effect (0.67).

5 Discussion

Our cases studies highlight several benefits of incorporating contextual factors into social media analyses. First, they support disambiguation. In the first case study, two pseudo-cliques for locations emerged from the data derived from geo-coded Twitter posts. Given their overlap in terms of the locations themselves, the contextual factors shown in Figure 8 provide clues: The clique we labeled 'New York Tourist' because the members are predominantly not from the NYC

[5] As the pair-counting F-measure is not a symmetric measure between two clustering assignments, we report the mean of the two directional comparisons.

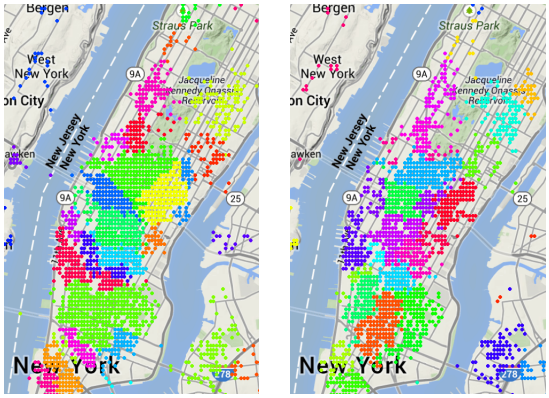

Figure 12: Neighborhoods based on gender (male on left, female on right)

metropolitan area is more gender balanced and reflects more positive mood than the 'Midtown Worker' clique. The contextual factors provide validation for the difference between the two cliques. This means that locations in Manhattan are clustered together in different graphs depending on the type of person you are, which in turn is reflected in the contextual factors of gender, metro area, and mood.

Second, the discussion graph framework allows the combination of the strengths of social media data (geo-coding, time-stamps) with more traditional person variables like every day behavioral patterns (weekday/weekend) and gender in order to develop more nuanced interpretations of results. As we saw in the second case study, neighborhood boundaries shifted substantially when conditioned on these different factors, even when using the identical boundary drawing algorithm. In Figure 11 we showed that shopping districts emerged on the weekends. Figure 12 shows substantial neighborhood differences for men and women, with men splitting apart the neighborhood on the west side of lower central park, but maintaining much larger neighborhoods south of the park than do women. That is, the behavioral patterns of men and women may be sufficiently different to suggest different neighborhoods. These distinctions suggest extensions to lines of research. In this case, we can suggest that not only can neighborhoods be defined by behavior patterns, as in LiveHoods, but can be further refined by contextual factors such as day of the week and gender.

6 Conclusions

Motivated by the importance of understanding how the context of social media discussions affects the information we extract from them, we designed and built a framework that simplifies the specification and work required to deeply explore and condition results on context. To this end, we presented discussion graphs, a data model for co-occurrence analyses, and DGT , our implementation. This data model and implementation capture the computations and data representations common to co-occurence analyses of social media data across many domains, and jointly represent the structure and the context of relationships inferred from social media. Our goal is to take a step toward enabling the type of conditional analyses typically employed in the social sciences, while taking advantage of the unique properties of

social media data. Supporting these types of analyses will allow us to understand important distinctions such as how gender, time, and mood affect results, rather than simply averaging over conditioning factors.

Through several examples and case studies, we show how the discussion graph framework greatly simplifies the task of building a social media analysis. This improved agility has significant implications for social media research: we demonstrate that high-level analysis results, such as inferred neighborhood boundaries, depend on the context from which social relationships were extracted,illustrating that deeper analysis is necessary for truly understanding the information we extract from social media.

We expect to make our core system and feature extractors available publicly soon, with the goal of spurring broader and more complete investigations 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.

References

Bastian, M.; Heymann, S.; and Jacomy, M. 2009. Gephi: an open source software for exploring and manipulating networks. In *Proceedings of ICWSM 2009*.

Bollen, J.; Mao, H.; and Zeng, X. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2(1):1–8.

Calais Guerra, P.; Meira, W.; and Cardie, C. 2014. Sentiment analysis on evolving social streams: How self-report imbalances can help. In *Proceedings of WSDM 2014*. ACM.

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

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013a. Major life changes and behavioral markers in social media: case of childbirth. In *Proceedings of the 2013 conference on Computer supported cooperative work*, 1431–1442. ACM.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013b. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of CHI 2013*, 3267–3276. ACM.

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

De Longueville, B.; Smith, R. S.; and Luraschi, G. 2009. Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 Intl. Workshop on Location Based Social Networks*, 73–80. ACM.

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

Golder, S. A., and Macy, M. W. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051):1878–1881.

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

Hansen, D.; Shneiderman, B.; and Smith, M. A. 2010. *Analyzing social media networks with NodeXL: Insights from a connected world*. Morgan Kaufmann.

Heer, J., and Perer, A. 2011. Orion: A system for modeling, transformation and visualization of multidimensional heterogeneous networks. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, 51–60. IEEE.

Kıcıman, E. 2012. OMG, i have to tweet that! a study of factors that influence tweet rates. In *Proceedings of ICWSM 2012*.

Konstas, I.; Stathopoulos, V.; and Jose, J. M. 2009. On social networks and collaborative recommendation. In *Proceedings of SIGIR*. ACM.

Kosinski, M.; Stillwell, D.; and Graepel, T. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110(15):5802–5805.

Leskovec, J. 2013. Stanford network analysis package (snap). `http://snap.stanford.edu/`. Accessed: 2014-03-24.

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

Liu, Z.; Navathe, S. B.; and Stasko, J. T. 2011. Network-based visual analysis of tabular data. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, 41–50. IEEE.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. N. 2011. Understanding the demographics of twitter users. In *Proceedings of ICWSM 2011*.

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

Myslín, M.; Zhu, S.-H.; Chapman, W.; and Conway, M. 2013. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research* 15(8).

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

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

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of WWW 2010*, 851–860. ACM.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Agrawal, M.; Park, G. J.; Lakshmikanth, S. K.; Jha, S.; Seligman, M. E.; Ungar, L.; et al. 2013. Characterizing geographic variation in well-being using tweets. In *ICWSM'13*.

Weng, J.; Lim, E.-P.; Jiang, J.; and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of WSDM 2010*. ACM.