Using Spectral Clustering of Hashtag Adoptions to Find Interest-Based Communities

Aurora Schmidt and Clay Fink
Johns Hopkins University
Applied Physics Laboratory
11100 Johns Hopkins Rd, Laurel, Maryland
{aurora.schmidt,clayton.fink}@jhuapl.edu

Vladimir Barash
Graphika, Inc.
116W 23rd St. New York, NY
vlad.barash@graphika.com

Christopher Cameron and Michael Macy
Cornell University
Social Dynamics Laboratory
372 Uris Hall, Ithaca, NY 14853
{cjc73,mwm14}@cornell.edu

***Abstract*—We investigate the use of spectral clustering of hashtag adoptions in Nigerian Twitter users between October 2013 and November 2014. This period is of interest due to the online campaign centered around the #BringBackOurGirls (BBOG) hashtag, which relates to the kidnapping of 276 Nigerian schoolgirls. We examine the adoption of hashtags during the six months before, the month after, and the six months following the kidnapping to test the informational value of behavior-based clusters discovered with unsupervised methods for predicting future hashtag usage behaviors. We demonstrate an efficient spectral clustering approach, that leverages power iteration on symmetric adjacency matrices, to group users based on hashtag adoptions prior to the kidnapping. Unlike follow network-based clusters, these adoption-based clusters reveal groups of users with similar interests and prove to be more predictive of interest in future topics. We compare this unsupervised spectral clustering to spectral clustering based on symmetrized follow network relations as well as clusters induced by latent Dirichlet allocation (LDA) topics. We find that hashtag adoption-based clusters perform similarly to the more computationally expensive LDA approach at identifying interest groups that are more likely to adopt future topical tags. We also benchmark the spectral clustering approach against the popular Louvain clustering approach on a synthetic dataset, finding the faster spectral clustering algorithm produces more balanced clusters with a higher similarity to the true interest groupings used to synthesize adoption data.**

I. Introduction

The opportunities presented by advances in graph signal processing and use of the graph spectra to analyze similarity and adjacency matrices have great potential to contribute to the analysis of large datasets generated by online social media populations. We propose a method of spectral clustering that efficiently finds communities of users with similar interests based on hashtag adoption behaviors. The clustering of users into interest-based communities provides insight into how prior interests in topical areas can affect the adoption and spread of new movements or campaigns in online communities. By finding user clusters that are predictive of individual susceptibility to various topics, we enable that information to be incorporated into agent-based contagion models that predict the impact and spread of messages in social networks.

The adoption of topical hashtags by communities with low levels of prior interest in a topic can also indicate an unusually strong impact of a newly trending discussion or social movement; a phenomenon known as "breakout". This indicator is of interest to computational social scientists that study the factors leading to the spread of social movements in populations. Theories of complex contagion predict that behaviors such as joining protests or expressing controversial views often require social reinforcement among peers, and, as such, network level markers of the spread of opinions that are risky may include the achievement of a critical mass of densely connected early adopters followed by breakout of adoptions into new communities, [1], [2], [3]. One of the goals of categorizing users by prior indications of topical interests is to detect the spread of a political or controversial tag beyond a small number of communities, which can, along with other factors, indicate a sea change in expressed attitudes and opinions in the population.

We detail an unsupervised process for discovering clusters of individuals that use similar hashtags and show how we may gain insight into the interests of the groups by examining tags that differentiate them from the rest of the population. We show that detected cluster membership explains correlation in the rates of adoption of new tags and an increase in the potential information gain available to an optimal predictor of future tag adoptions. We compare that clustering method to two alternate approaches based on follow network ties as well as Latent Dirichlet Analysis (LDA) for topic discovery. We show that spectral clustering with only user data regarding adoption of hashtags often outperforms spectral clustering of follow network ties for potential information gain in future topical tag adoptions. We also show that spectral clustering with just adoption data can approximate the more data- and computation-intensive approach of LDA, leading to very efficient methods for group discovery and characterization. Lastly, we compare spectral clustering to the popular Louvain clustering method, [4], on a similarity matrix derived from synthetic adoptions in which true interest categories are known, finding that spectral clustering produces more balanced clusters that better match the true interest groups while running 14 times faster than the popular python implementation of the Louvain algorithm.

The following section gives background and related work. The Data section describes the data set collected for analysis of Nigerian Twitter users. Methods provides the clustering

978-1-5386-3180-5/18/$31.00 ©2018 IEEE

algorithms used, the approach to examine the distinguishing hashtags of detected clusters, the measures used to estimate the predictive power of the clusters regarding adoptions of future hashtags. and the measures of similarity between multiple clusterings. In Experiments and Results, we present the unsupervised clusters detected by the three compared clustering methods. We show the top distinguishing tags for selected groups found based on adoption patterns. We further compute how effectively users are sorted into groups that differentially reacted to post-kidnapping tags as well as correlations in the adoptions of topically relevant tags over the adoption-based user clusters. Finally, we compare the proposed approach to spectral clustering with Louvain clustering on a synthetic dataset.

II. Background and Related Work

The idea that social media could facilitate protest, revolution, and social change was initially met with skepticism, especially regarding its use during the protests that followed the 2009 Iranian election [5], [6]. Subsequent empirical work has since documented the role of online activism in social movements; e.g., the Iranian elections [7] and 2011 Egyptian revolution [8]. Research on the #BringBackOurGirls movement includes [9], which examined differing topologies of Twitter user networks using the hashtag, identifying "broadcast networks" of high profile accounts and "community clusters" of lower profile activists that kept the movement alive by sharing new information and coordinating offline activism.

Our work focuses on unsupervised methods to find communities with similar interests, examining how well the discovered community membership can predict future interests. Because we use past user behavior to find groups that react to similarly to topics labeled by usage of hashtags, we also obtain taxonomies of hashtags that distinguish various user groups. Prior work in this area includes [10], [11], [12], [13]. Our work is most similar to that of [10], which used a method that combined topical interests as well as social ties into an expanded matrix which they used to perform spectral clustering on the graph Laplacian. Using the spectrum of adjacency matrices is gaining popularity in applications of graph signal processing; see [14], and has noted equivalences to Laplacian methods, [15]. An overview of spectral clustering is provided in [16]. In [10], the problem sizes were smaller, using approximately 3,000 users and 3,000 followed blogs containing tag labels. Our dataset contains over 20,000 users and over 70,000 hashtags used to cluster the users. As a result, we adopt a more computationally efficient approach to spectral clustering using the weighted adjacency matrix computed based on similarity in the vector space of user hashtag adoptions. Because we use the similarity matrix, we may focus on finding the largest eigenvectors with which to perform clustering. This allows for use of the more scalable method of the power iteration for computing eigenvectors, as opposed to more intensive matrix eigendecomposition techniques.

Our objectives are very similar to those in [13]. However, Shi and Macy performed clustering using an alternate metric for similarity, called standardized co-incident ratio, showing lower sensitivity to wide variations in out-degree of category nodes. This superior similarity metric may be used with our technique instead of cosine similarity, but requires the computation of the expected number of shared hashtag adoptions between users in a randomized adoption network. We leave a study of improved similarity metrics combined with spectral clustering to future work and focus this work on the formation of clusters from similarity matrices.

In [11], the authors rely on entity detection and disambiguation as well as a topic folksonomy harvested from Wikipedia. We, instead, rely on the user generated hashtags to differentiate the topical content of tweets. The authors of [12] also treat the problem of user interest detection. They use Twitter lists, which users may choose to follow, as an indicator of interest and find very high agreement between the detected topics of lists to the profile descriptions that users provide of self-stated interests. Using hashtags to study user interests has the advantages that they are heavily used and are created very quickly in response to emerging news and social discourse. In addition, our metric for evaluating the performance of our detected clusters of similarly interested users differs from [12]; we measure how well clusterings of users predict each group's adoption rates of new tags. In essence, we seek a clustering of users that can yield predictive insight into how the users will react to new events.

III. Data

We collected tweets from public Twitter accounts via the free API using geofenced search queries across Nigeria. Over a six year period, we monitored 45 bounding circles placed to cover cities with populations of over 100,000. Geofenced search results include tweets that are geotagged or have profile locations within the requested bounding circle. From the returned tweets, we randomly select a sample of the most recently active users and obtained public timeline histories that include up to the last 5000 past tweets. For years 2013 through 2015 the data set contains 434 million tweets from 3.2 million users. The data was stored in a MongoDB database and indexed using Lucene to facilitate text searches. Data collection was conducted under an IRB exemption and is stored with anonymized Twitter user and tweet identifiers to protect user privacy.

For this paper, our analysis period covers October 1, 2013 through October 31, 2014, a 13 month period centered around the Chibok schoolgirls kidnapping which occurred on April 15, 2014. We split the time period into the six months before the kidnapping (October 1, 2013 through April 14, 2014), the month immediately following the kidnapping (April 15, 2014 through May 14, 2014), and the six months following (May 15, 2014 through October 31, 2014). Since we were interested in behavior across the full analysis period, we identified users that had recorded twitter activity in our database in all three sub-periods and had accounts created before October 1, 2013, providing a set of 197,022 users.

Follow links for users who had been active during 2014 in our larger Nigerian data set were collected using the Twitter API.[1] This included follow data for 24,672 of the 197,022 identified users. Another clustering approach was based on the topic distributions of users derived from running Latent Dirichlet Allocation (LDA), where we treated the concatenation of each user's tweets as a single document. To give us documents of adequate size for generating the topic models we restricted our data to the set of users who had 25 or more tweets in the first period of analysis. This left us with subset of 22,293 of the 24,672 users with follow graph data. From these users we had a total of over 42 million tweets across the entire time period. The results in this paper are based on this set of 22,293 users.

We then selected hashtags that were used during the analysis period 50 or more times. This set of 125,485 tags was used to create user-hashtag adoption matrices for each of the three periods. The adoption matrix for the first period was restricted to just the 71,661 hashtags used before April 15, 2014. Those from the later two periods were used for to assess similarity and predictability of hashtag adoptions based on clustering during the period before April 15th.

IV. Methods

In this section, we detail the spectral clustering algorithm used as well as the approach for finding the distinguishing hashtags of each detected group of users. We also define the methods for comparing clustering outputs and measuring the predictive power of the clusters regarding future hashtags adoptions.

A. Unsupervised Spectral Clustering

Spectral clustering relies on the eigendecomposition of the matrix defining ties between the nodes of a graph. We model the users as nodes of a graph and use one of two options for an adjacency matrix. The first option is a thresholded similarity metric on prior hashtag adoptions of the users. The second option, for comparison, uses a symmetrized version of follow network relations. Both choices result in weighted symmetric adjacency models, and are convenient due to the guarantees of existence of a complete set of distinct eigenvalues and orthonormal eigenvectors. We first describe the construction of a graph based on similarities in user hashtags adotions.

For the graph of users, $\mathcal{G} \equiv (\mathcal{V}, \mathcal{E})$, we model as vertices the set of users, and as edges the real values measuring the similarity in the set of hashtags each pair of users tweeted with during the training data period. For the set of all tags that had at least 50 uses, we create a vector for each user, denoted $\mathbf{x}_i \in \mathbb{Z}^p$, that assigns a 0 to all tags not used and a 1 to all tags used by the i^{th} user. These vectors corresponding to the adoption pattern of each user is collected into an adoption matrix, P, given by

$$P \in \mathbb{Z}^{n \times p} \quad = \quad \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \tag{1}$$

In our data set, the number of users n was 22,293 and the number of hashtags was 71,661. Due to the size and sparsity of the matrix, the spectral clustering algorithms were implemented using sparse matrix structures provided by MATLAB R2015a and in Python using the Compressed Sparse Column matrix type in the Scipy package. The maximum number of hashtags used by a single user was 2,687 and the minimum was 1. We compute the similarity matrix using the cosine similarity between pairs of adoption vectors, resulting in the similarity matrix, S. Defining $\tilde{P}$ as the adoption matrix with rows scaled to have unit norm, the similarity matrix is given by

$$S \in \mathbb{R}^{n \times n} \quad = \quad \tilde{P}\tilde{P}^T - I \tag{2}$$

where I is the n-by-n identity matrix, removing all self ties that would appear on the diagonal. For efficiency, we threshold these innerproducts to store only similarities greater than or equal to 0.1 resulting in 20.5 million values and a density of 4.1% nonzeros. These similarity values measure the normalized overlap in hashtag adoptions between each pair of users. The network of users formed by this similarity matrix would lead to random walks in which users with similar interests are the most likely to be visited.

A clustering based on the graph spectra; i.e., eigenvectors, of S would attempt to partition this graph of users in a way that preserves the dominant features of the matrix, corresponding to the eigenvectors associated with the largest eigenvalues of S. That is, for the eigendecomposition of the symmetric weighted adjacency matrix, $S = U\Lambda U^T$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $\lambda_1 \geq \ldots \geq \lambda_n$, and $S\mathbf{u}_i = \lambda_i \mathbf{u}_i$ for every column of U, we will use the top k eigenvectors in U as lower dimensional coordinates for clustering of users. We will use 50 as the chosen number of clusters for analyses of the 22,293 users. The algorithm in fig. 1 denotes this approach of selecting the top k eigenvectors, and using the k-means algorithm to generate k clustered groups. This approach is most similar to that of Shi and Malik, summarized in [16], except that it uses the graph adjacency rather than the Laplacian.

For very large matrices, we prefer to perform spectral clustering with the weighted adjacency matrix. This is because the eigenvectors used correspond to the largest eigenvalues, which can be found much more efficiently that fully diagonalizing the

```
function SPECTRALCLUSTERS(A, k)
    (U_k, Λ_k) ← TOPEIGS(A, k)
    (C_1, C_2, ..., C_k) ← kmeans(U_k, k)
    return (C_1, C_2, ..., C_k)
end function
```

Fig. 1. Perform spectral clustering on network or similarity matrix A

[1]The follow data was retrieved in June of 2015 so some links between the users may not have been present during the analysis periods. The API does not return time stamps for when a follow relationship was created.

function EIG($A, \Lambda_{m-1}, U_{m-1}$)
  $n \leftarrow \text{size}(A, 2)$
  $x \leftarrow \text{rand}(n, 1)$
  $x \leftarrow x - U_{m-1} \,\text{diag}\left(x^T U_{m-1}\right)\mathbf{1}_{m-1}^T$
  $x \leftarrow x/\|x\|$
  for $i \leftarrow 1$ to 1000 **do**
    $x' \leftarrow Ax$
    $x' \leftarrow x' - U_{m-1} \,\text{diag}\left(x^T U_{m-1}\right)\mathbf{1}_{m-1}^T$
    $\lambda \leftarrow \|x'\|$
    $x' \leftarrow x'/\lambda$
    if $i > 100 \wedge \left(1 - x^T x'\right) < \epsilon$ **then break**
    end if
    $x \leftarrow x'$
  end for
  return (x, λ)
end function

Fig. 2. Power iteration to find m^{th} largest eigenvector of A, given the previously found Λ_{m-1} and U_{m-1}

matrix, S. Algorithm 2 shows how the power method is used to compute the m^{th} largest eigenvalue-eigenvector pair, given the previously computed $m-1$ eigenvalues and eigenvectors. The method relies on the fact that after repeated multiplication of a random initial vector by a matrix, the vector will converge to that corresponding to the largest eigenvalue of that matrix. By removing the projection onto the previously computed eigenvectors, which is simple due to the orthogonality of the eigenvectors of a symmetric A, we can sequentially compute the top k eigenvectors needed for spectral clustering.

To compare our method of clustering using hashtag adoption patterns, we also perform spectral clustering using follow information on the set of 22,293 users. For this we construct a follow graph $F \in \mathbb{Z}^{n \times n}$, where there is a 1 placed at each (i, j) location where user i follows user j. This social relation is not a symmetric one; that is, i following j does not imply that j follows i. To ensure existence of the eigendecomposition of the network matrix, we project this follow matrix to a symmetric one

$$A \in \mathbb{R}^{n \times n} \quad = \quad \frac{1}{2}F + \frac{1}{2}F^T \tag{3}$$

There were almost 1.8 million follow relationships between users in the set, resulting in a sparse A matrix with a density of about 0.5% nonzeros. Spectral clustering results using the similarity of hashtag adoptions, given by S, are compared with clustering using the symmetrized follow network, A, to benchmark the approach and to test whether clustering with hashtag usage provides better predictions of future tag adoptions than that of the follow network. In addition, we compare our method on a clustering of the document vectors from a 100 topic LDA model from a set of documents where each document contains total tweets from one of the 22,293 users. Each document vector captures the distributions of 100 topics for a given user. We then use k-means to obtain 50 clusters of topically similar users. Lastly, we compare spectral clustering to Louvain clustering on a synthetic adoption matrix for 2000 users, testing similarity of each clustering result to the known interest groups used to generate the adoption data.

B. Comparing Similarity of Clustering Results

We use two measures to compare the similarity of outputs of clustering approaches. The first computes the likelihood that two methods will agree on the co-membership status of a randomly selected pair of users. This likelihood is computed by counting the number of pairs that exist in method A but not in method B and vice versa. Since these encompass the only types of errors, we then normalize the sum of those errors by the total number of distinct pairs, $n(n-1)/2$. We compare that to the likelihood that two cluster results that are obtained through random assignment; i.e., choosing a user's cluster uniformly at random from 1 to 50. For two random assignments the likelihood of disagreement is $(1/50)(49/50) + (49/50) * (1/50) = 0.039$.

The second method we use to compare clustering results is that proposed by [17]. For two sets of clusters, $C = \{C_1, \ldots, C_m\}$ and $D = \{D_1, \ldots, D_n\}$ the similarity is given by computing a similarity matrix,

$$S_{C,D} \quad = \quad \begin{bmatrix} S_{11} & \cdots & S_{1n} \\ \vdots & \ddots & \vdots \\ S_{m1} & \cdots & S_{mn} \end{bmatrix} \tag{4}$$

where $S_{i,j}$ is given by the ratio of the size of the intersection of common members in C_i and D_j divided by the union of the unique members in the two clusters. Then the overall similarity of the two clustering results is

$$Sim(C, D) \quad = \quad \sum_{i=1}^{m}\sum_{j=1}^{n} S_{ij} / \max(m, n) \tag{5}$$

With this measure the most similar result would yield a value of 1, implying identical clusters.

C. Distinguishing Hashtags

After finding clusters of users displaying similar interests, we would like to automatically list the tags that each group disproportionately adopted. This helps us to gain insight into what makes a user set unique. For example, if the vast majority of the overall population uses a common hashtag like #tbt, for throw-back-thursday, we do not want to list that as a distinguishing hashtag for a group that also frequently uses that tag. Our approach is to select the top r tags that meet two requirements: 1) the hashtag must be adopted by greater than 1% of the members in the group and 2) they are the r such hashtags that have the highest binomial tail values, denoted b. By binomial tail value, we mean that if the fraction of adopters of the tag, computed over all users, is p, and the number adopting in the group of size s is N, the binomial tail value is the cumulative probability of observing less than N adoptions out of s trials with success probability, p. This method of examining group-specific tags helped us find topically interesting tags that were perhaps only adopted

within a local cluster of similarly interested individuals. We also hand-labeled a set of topically relevant popular tags which we use in the next section to measuring information gained by cluster membership in the likelihood of adoption of future topical tags.

V. Experiments and Results

In this section, we present the unsupervised clusters detected by the four compared clustering approaches on the Twitter dataset of 22,293 users and on a synthetic dataset of 2000 users with known interest groups. We show the top distinguishing tags for selected groups found based on adoption data. We further compute how effectively users are sorted into groups that differentially react to post-kidnapping tags as well as correlations the adoptions of topically relevant tags over the adoption-based detected user clusters.

A. Unsupervised Clustering of the Twitter Dataset

We computed a clustering over 22,293 users using the 3 clustering methods described in section IV. We compare spectral clustering with the similarity matrix, computed using the adoption matrix for the first analysis period, spectral clustering based on a symmetric projection of the follow graph, and clustering based on LDA document vectors. We used the Mallet toolkit [18] to generate a topic model from the user documents based on 100 topics, which took approximately five hours to run on a four core system with 16 GB of RAM. In comparison our MATLAB tools for spectral clustering completed in less than 30 minutes on a similar system.

We find that the spectral clustering method produces clusters with more balanced sizes. Since all 3 methods were configured to generate 50 clusters, they all have a mean membership size of 445.9. Adoption-based, network-based, and LDA-based clusters had a standard deviation in membership of 137.6, 170.0, and 352.0, respectively. The LDA method had dramatically less balanced clusters, potentially giving it an advantage in predicting the distributions of adoption of tags in future periods due to having more users as members of larger groups.

We compared the similarity of these clustering results by the two metrics detailed in Methods. For the following we will refer to adoption based clusters as A, network based as B, and LDA based as C. The fraction of mismatched pairings between the three methods were (AB, 4.3%) , (AC, 5.0%), and (BC, 5.1%). All of these were greater than the expected mismatch rate of 4.3% that would be observed with random cluster assignments, indicating that each method was arriving at quite different groupings. By the similarity metric of Torres et al., given in IV-B, the similarities measured were (AB, 0.48), (AC, 0.44), and (BC, 0.44%), also indicating low similarity between the results of each method.

In IV-C, we describe the selection of distinguishing hashtags for each group. Table I shows the tags that distinguished 3 clusters found by adoption based clustering, showing the most distinctive tags for periods 1, 2, and three on the upper middle and lower rows, respectively. Cluster size is shown below the

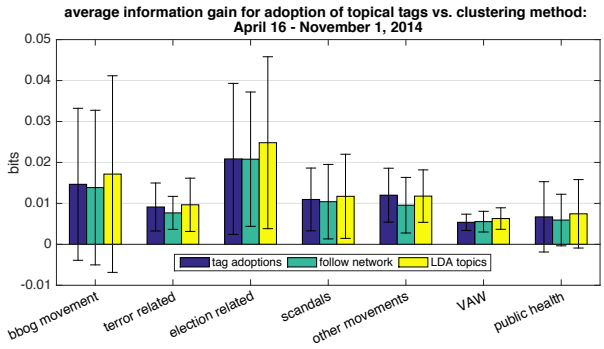

Fig. 3. Average information gained due to cluster membership on future topical hashtag adoption rates.

cluster number. These clusters were presented due to their strong participation in social and political discussion. Many of the other 47 clusters were primarily interested in football, pop culture, or religious matters. Group 27 is very engaged with terror events, group 29 is both engaged with news and politics using a number of hashtags indicating interest in the opposition party, APC, and its leader, Buhari, and cluster 50 found users that were very engaged in political discourse through the Young African Leadership Initiative network. In addition, we see these politically active groups are likely to continue discussion of similar topics in periods 2 and 3, indicating that clustering based on period 1 adoptions can group members who are likely to continue with similar topical interests.

We hand selected the tags that corresponded to 7 topics of interest from the top 30 distinguishing tags of the 50 groups during the 3 periods. These tags are shown in Table II. To measure the predictive power of the compared clustering methods, we computed the information gain due to knowing cluster membership to the distribution of adoption rates of a randomly selected user. We then show the average and standard deviations of information gains of each cluster type in Figure 3. This was measured by the mutual information between cluster ID and adoption. We see that LDA clustering generally resulted in more information gained, and adoption based clusters were slightly better than follow network clusters. However, in general, all 3 techniques averaged similar gains with large variations per topic.

We also computed the predictive power of each set of clusters on the level of engagement of members with BBOG tags. We looked at the total number of distinct days of tag usage of BBOG as well as the duration; i.e., the difference between the time of latest and earliest use. The information gained using adoption-, network-, and LDA-based clusters to predict the number of days users referenced BBOG tags was 0.1830, 0.1607, and 0.2025 bits, respectively. Again LDA measured the highest gain, which could be related to having less balanced cluster sizes, and adoption-based clusters improved on network clusters. The respective information gained regarding total duration of BBOG engagement was 0.4076, 0.3910, and 0.3717 bits, showing with adoption clusters outperforming both network and LDA clusters.

27 (408)	nigeria bokoharam abuja borno abujabombblast nyanya nyanyablast nyanyabombblast nyanyaupdate bombblast nigeria bringbackourgirls chibokgirls abuja chibok bokoharam wefafrica bringbackourdaughters nigerian nyanya nigeria bringbackourgirls chibokgirls bokoharam abuja chibok humanrightstulip2014 nigerian ekitidecides ebola
29 (474)	nigeria bokoharam change apc buhari gej pdp breakingnews nigerian vanguardngrnews nigeria bringbackourgirls chibokgirls bokoharam abuja apc gej pdp nigerian gmb nigeria bringbackourgirls chibokgirls bokoharam abuja ebola change apc nigeria2015 buhari
50 (386)	africa in yalichat entrepreneurship socent yali yali2014 socialentrepreneurship leadership business nis yalichat yali youth yalinetwork usafrica entrepreneurship entrepreneur africa leadership impact2014 yalichat yali2014 washingtonfellows yali mentalhealthmatters youth sw4ag11 stopcsobill yali2015 socialgood

TABLE I
TOP 10 DISTINGUISHING HASHTAGS OF 3 SELECTED CLUSTERS DETECTED BY SPECTRAL CLUSTERING BASED ON HASHTAG ADOPTION SIMILARITY. THE LINES CORRESPOND TO PERIODS 1, 2, AND 3, WITH THE TOP LINE (PER. 1) CORRESPONDING TO BEFORE THE CHIBOK KIDNAPPING.

BBOG Movement	bringbackourgirls whereareour85daughters whereareourdaughters chibokgirls chibok chibok234 bringbackourdaughters bringbackoursisters bringourgirlsback bringbackthegirls bringback bbog whereareourgirls missing bornogirls prayforborno abductedbornogirls saveourgirls 234girls chibokwomanleader yaallahsaveourgirls bringbackoursgirlsnowandalive bringourgirlshome chibokwomen bringbackourgir releaseourgirls bringbackour 200missinggirls thechibokgirls bringabackourgirls bringbackmylittlesisters schoolgirls notwithoutourdaughters nigerianschoolgirls prayforpeace freeourgirls
Terror Related	bornogirls prayforborno bornomassacre nyanyabombblast nyanyablast notobokoharam endbokoharam stopkidnappingbyyinqueafrique bringbackthepeace gejpleaseaskforhelp endterrorism bokoharamisnotislam bringbackourcountry citizenssolutiontoendterrorism stopthebombings stopthebombing abujabombblast nyanyablast2 stopbokoharam nyanyabombblastagain
Election Related	election2015 kwankwasiyyaamana rmk osundecides ekitidecides bringbackjonathan2015 bringbackgoodluck2015 bringbackjonathan gej2015 teambuhari idreamofanigeria victoryfornigeria gmb gejmustgo stepdownjonathan gejstepdown pdp apc gej gejmediachat presidentialmediachat
Scandals	freeciaxon whereisourmoney sanusi nis nismurder nisexam freenaominyadar americawillknow
Other Movements	stolendreams may1st workersday beijing20 childmarriage occupynigeria 30percentornothing shapeafrica2014 audacity2lead yalichat yali yali2014 yali2015 weareallmonkeys saynotoracism thebudgetisaroundyou climatechange timetoact enoughisenough
Violence Against Women	childmarriage aartalk choice4life date360 domesticviolence endchildmarriage endfgm endmaternaldeaths endviolence fgm fightrapethursday girlsummit give2ster jadapose justiceforjada rape saynotorape sexualviolence standtoendrape stopdomesticviolence stoptheviolence vappbill vaw whyistayed yesallwomen
Public Health	ebola factsonebola ebolatips ebolafacts ebolaoutbreak ebolavirus stopebola ebolanews ebolaresponse ebolachat ebolaqna ebolathemovie ebolaalert ebolafree ebolawatch askebola ebolafreenigeria ebolafact ebolafreeph ebolascare fearofebola ebolawatchng polio endpolio worldmalariaday malariafreenigeria defeatmalaria breastcancerawareness aids2014 worldsicklecellday

TABLE II
TOPICAL HASHTAGS WITH 50 OR MORE USERS BETWEEN APRIL 15 AND NOVEMBER 1, 2014

Cluster ρ	Indiv. ρ	Tag 1	Tag 2
-0.526	-0.014	malariafreenige-ria	ebolafacts
-0.355	-0.010	bringbackthep-eace	chibok
0.993	0.602	pdp	apc
0.967	0.584	timetoact	sexualviolence
0.960	0.144	releaseourgirls	notobokoharam
0.952	0.625	choice4life	vappbill
0.949	0.137	whereisour-money	gejpleaseask-forhelp

TABLE III
CLUSTER-BASED PEARSON CORRELATIONS, USING SPECTRAL CLUSTERS BASED ON PREVIOUS HASHTAG ADOPTIONS, OF PAIRS OF TAGS IN THE PERIOD AFTER THE KIDNAPPING

We used the adoption-based clusters to measure correlations between group adoption rates of pairs of tags that had at least 50 adoptions in the periods after the kidnapping. We also compared the measured cluster-based correlation to the measured correlations over the adoption of individuals. In general, the cluster correlations were much higher, indicating value to summarizing the likelihoods of adoption by membership within interest-based groups. We look at both negatively and positively correlated hashtags, with selected top tags presented in Table III. The negative correlation between #MalariaFreeNigeria and #EbolaFacts may indicate that attention for the two health issues is competitive. The tag #BringBackThePeace had different usage statistics than other BBOG-related tags, perhaps indicating a different message that is less focused on the abduction. The two tags #PDP and #APC are the acronyms of the two major political parties in Nigeria; their high correlation could indicate that adoption of the tags does not give good predictions of party affiliation or that our clusters may be grouping members of both parties together. Given the high correlation of #PDP and #APC on the individual level, it looks more like politically engaged individuals are likely to use both tags. The correlations between #BringBackOurGirls and

#OccupyNigeria as well as #BringOurGirlsBack and #StolenDreams are interesting as #OccupyNigeria and #StolenDreams refer to organized protests, and so we may be seeing the strong support by broad activists for the BBOG movement. There is much insight to be gained using unsupervised methods for extracting relationships between hashtags.

B. Comparison to Louvain Clustering With Synthetic Data

Method	Num. Clusters	Similarity to Truth	Wall Time
Louvain	5	0.1217	2.02 min
Spectral	10	0.2023	0.14 min

TABLE IV
COMPARISON OF SPECTRAL TO LOUVAIN CLUSTERING ON A SYNTHETIC DATASET WITH 2000 USERS, 5 TOPICS, 56 HASHTAGS, AND 31 DISTINCT TOPIC-BASED USER GROUPS

To create a synthetic dataset where true interest groupings are known, we created 5 synthetic topics. Each of these topics had a mean of 10, standard deviation of 8, tags. There were 56 total tags. For 2000 users, we sampled each individual's number of interests with a mean of 2 and standard deviation of 3[2]. Each user then randomly samples the topics corresponding to their interests from the uniform distribution of the 5 topics. This results in 31 groups of unique interest combinations, where there is a minimum of 20 users in the smallest group and 394 in the largest group with a mean membership size of 64.5 users. Each user chooses uniformly at random to adopt between 5 and 18 tags per topic and then samples these uniformly at random from the tags for each topic of interest. From this we obtain a sparse adoption matrix with 56 tags and 2000 users. We compared the time to compute 10 spectral clusters using a python implementation of our method to running Louvain with python community, which yielded 5 clusters, as presented in table IV. Our method took 0.14 minutes while Louvain took 2.02 minutes on a 4 core system with 16 GB of RAM. The Louvain method is designed to optimize modularity of the clusters and achieved modularity of 0.1803 while spectral clustering achieved 0.1066. However, the similarity of the spectral clusters to the true groups, by the Torres et al. metric, was 0.2023, while Louvain clusters had 0.1217 similarity to truth, with only 0.4601 similarity to the spectral clusters. Spectral clusters had more balanced sizes; the standard deviation in the membership of Louvain groups was 140.2 while for spectral clusters it was 41.4.

VI. CONCLUSIONS

We developed a scalable method of unsupervised detection of clusters based on patterns of Twitter hashtag adoptions using a form of spectral clustering. We compared this method to spectral clustering of the symmetrized follow network as well as clustering based on LDA topic distributions, finding that adoption-based spectral clustering was somewhat more predictive of future adoptions than follow graph clustering. LDA topic clustering outperformed both, but requires analysis of all language in tweet histories and is much more computationally taxing. In addition, we showed for a synthetic adoption dataset, spectral clustering yielded clusters with higher similarity to true user groups than Louvain clustering and was 14 times faster. In the future, we plan to investigate clustering approaches that combine both adoption and network information as a weighted combination to future improve predictive power of resulting clusters as well as test new similarity metrics.

[2]A minimum of 1 tag per topic and 1 interest per user was imposed.

ACKNOWLEDGMENT

This work was funded by the Minerva Initiative through the United States Air Force Office of Scientific Research (AFOSR) under grant FA9550-15-1-0036.

REFERENCES

[1] V. Barash, C. Cameron, and M. Macy, "Critical phenomena in complex contagions," *Social Networks*, vol. 34, no. 4, pp. 451 – 461, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378873312000111
[2] D. Centola and M. Macy, "Complex contagions and the weakness of long ties1," *American Journal of Sociology*, vol. 113, no. 3, pp. 702–734, 2007.
[3] M. S. Granovetter, "The strength of weak ties," *American journal of sociology*, pp. 1360–1380, 1973.
[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, Oct 2008.
[5] E. Morozov, "Iran: Downside to the" twitter revolution"," *Dissent*, vol. 56, no. 4, pp. 10–14, 2009.
[6] M. Gladwell, "Why the revolution will not be tweeted." *The New Yorker*, 2010.
[7] S. Aday, H. Farrell, M. Lynch, J. Sides, J. Kelly, and E. Zuckerman, "Blogs and bullets: New media in contentious politics," *United States Institute of Peace*, no. 65, 2010.
[8] Z. Tufekci and C. Wilson, "Social media and the decision to participate in political protest: Observations from tahrir square," *Journal of Communication*, vol. 62, no. 2, pp. 363–379, 2012.
[9] C. Carter Olson, "# bringbackourgirls: digital communities supporting real-world change and influencing mainstream media agendas," *Feminist Media Studies*, pp. 1–16, 2016.
[10] A. Java, A. Joshi, and T. Finin, "Detecting communities via simultaneous clustering of graphs and folksonomies," in *Proceedings of the tenth workshop on Web mining and Web usage analysis (WebKDD)*. ACM, 2008.
[11] M. Michelson and S. A. Macskassy, "Discovering users' topics of interest on twitter: A first look," in *Proceedings of the 4th Workshop on Analytics for Noisy Unstructured Text Data*, ser. AND '10. New York, NY, USA: ACM, 2010, pp. 73–80. [Online]. Available: http://doi.acm.org/10.1145/1871840.1871852
[12] P. Bhattacharya, M. B. Zafar, N. Ganguly, S. Ghosh, and K. P. Gummadi, "Inferring user interests in the twitter social network," in *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 2014, pp. 357–360.
[13] Y. Shi and M. Macy, "Measuring structural similarity in large online networks," *Social Science Research*, vol. 59, pp. 97–106, 2016.
[14] A. Sandryhaila and J. M. F. Moura, "Discrete signal processing on graphs: Frequency analysis," *IEEE Transactions on Signal Processing*, vol. 62, no. 12, pp. 3042–3054, June 2014.
[15] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," *Proceedings of the 2005 SIAM International Conference on Data Mining*, pp. 606–610, 2005.
[16] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.
[17] G. J. Torres, R. B. Basnet, A. H. Sung, S. Mukkamala, and B. M. Ribeiro, "A similarity measure for clustering and its applications," *Int. J. Electr. Comput. Syst. Eng.*, vol. 3, no. 3, pp. 164–170, 2009.
[18] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002. [Online]. Available: http://mallet.cs.umass.edu/