Bridging the Divide in Democratic Engagement: Studying Conversation Patterns in Advantaged and Disadvantaged Communities

Samah Gad*, Naren Ramakrishnan*, Keith N. Hampton† and Andrea Kavanaugh*
*Department of Computer Science
Virginia Tech, Blacksburg, VA, 24061
†School of Communication and Information
Rutgers University, New Brunswick, NJ 08901

***Abstract*—The Internet offers opportunities for informal deliberation, and civic and civil engagement. However, social inequalities have traditionally meant that some communities, where there is a concentration of poverty, are both less likely to exhibit these democratic behaviors and less likely to benefit from any additional boost as a result of technology use. We argue that some new technologies afford opportunities for communication that bridge this divide. Using temporal topic modeling, we compare informal conversational activity that takes place online in communities of high and low poverty. Our analysis is based on data collected through i-Neighbors, a community website that provides neighborhood discussion forums. To test our hypotheses, we designed a novel time series segmentation algorithm that is driven by topic dynamics. We embed an LDA algorithm in a segmentation strategy and develop an approach to compare and contrast the resulting topic models underlying time series segments. We examine the adoption of i-Neighbors by poverty level, and apply our algorithm to six neighborhoods (three economically advantaged and three economically disadvantaged) and evaluate differences in conversations for statistical significance. Our findings suggest that social technologies may afford opportunities for democratic engagement in contexts that are otherwise less likely to support opportunities for deliberation and participatory democracy.**

I. Introduction

Democratic engagement, at both the individual and community levels, is one of the strongest predictors of well-being [1]. While political behaviors, such as voting, are among the most studied aspects of democratic engagement, they are only a small subset of the behaviors that contribute to a democracy. Participation in a democracy involves more than the occasional selection of representatives. Citizens and their communities benefit from individual and collective action to address issues of common concern through activities outside of elections and government [2]. Participatory democracy includes a range of civic behaviors, including membership in institutions that address public issues, such as a neighborhood watch [3], as well as civil behaviors, such as helping a neighbor in an emergency [4]. These behaviors are intertwined with casual conversations, that, although not overtly deliberative or political, are a part of the "incomplete" [5] forms of political deliberation that are key to shaping social identities, friendships, and trust [6]. This combination of informal participation and casual, public deliberation provides for the social mixing that is important for opinion formation, awareness of common interests, social tolerance, and the ability to act on collective goals [7]. Unfortunately, like so many forms of democratic engagement, civic and civil behaviors and informal opportunities for deliberation are unequally distributed.

Civic and civil behaviors, including opportunities for informal deliberation, are stratified by class [8]. Those of lower income are significantly less likely to exhibit attitudes and behaviors for democratic engagement [2]. In addition, inequality is not equally distributed across the country, but concentrate in geographic areas of concentrated disadvantage; neighborhoods that are high in poverty, racial segregation, and social problems, such as crime [9]. The concentration of inequalities is associated with structural instability that reduce the ability of residents to form the local social bonds necessary for collective action [9]. As a result, those communities with the greatest need for informal discussion and participatory democracy are typically those where it is most absent.

Research on the role of new information and communication technologies ($ICTs$) and democratic engagement have generally found positive relationships between exposure to online political information and democratic behaviors [10], [11]. Participation in online activities that support informal deliberation, such as social networking services, has also been found to contribute to political participation [12]. However, there is almost no evidence that the use of ICTs overcomes existing socioeconomic inequalities associated with democratic engagement [13]. Indeed, there may be a "Matthew effect" [14], such that those who are already the mostly likely to express democratic behaviors gain further as a result of new ICTs, while those who have little gain little as a result of ICT use.

In this paper, we argue an alternative theory. We believe that new ICTs, specifically social media, offer new affordances for group interaction, informal deliberation and democratic engagement [15]. Unlike some other Internet technologies, social media afford contact in contexts where individuals have a shared affinity – through geography, political interests, or other interest – but previously lacked the means or ease of access for connectivity (in-person or online). In this paper, we focus on how these affordances reduce the cost of communication for urban communities with concentrated inequalities.

This reduction in the cost of communication helps residents overcome established structural barriers to social tie formation, informal deliberation and participatory democracy. The result is a set of opportunities for democratic engagement among people and in areas previously constrained by structural barriers to collective action. When such social media that are designed to bring local people together are made available to people in urban neighborhoods with high socioeconomic inequalities, we expect to find democratic engagement that is as high as what is typical of areas where such inequalities are less concentrated.

Specifically, our goal is to study the adoption of a tool for informal deliberation at the neighborhood level and to compare conversation patterns across advantaged and disadvantaged communities based on their level of concentrated poverty. Our aim is to characterize differences in informal deliberation, if any, between these advantaged and disadvantaged neighborhoods, as well as to detect common interests between them. This will provide insight into how neighborhoods with different poverty levels use ICTs for informal deliberation.

In order to be able to detect deliberation and common interests, we developed a novel temporal segmentation algorithm that is driven by topics discussed in a neighborhood setting. The objective of the algorithm is to detect segments where there are significant concordances of topics, but such that segment boundaries identify significant shifts in topics.

Once a neighborhood discussion is characterized in this manner, we can: compare the time duration of topics in neighborhoods with different poverty levels, identify differences in topics discussed between neighborhoods of different poverty levels, and identify differences in topics discussed between neighborhoods of similar poverty levels.

In the recent past, topic modeling techniques such as LDA (Latent Dirichlet Allocation; see [16] for a recent review) have emerged as a powerful approach to capture distributional trends in large text corpora. While classical LDA does not model temporal evolution, there are many variants of LDA [17], [18], [19] that do capture trends over time. Nevertheless, our needs here go beyond these systems, since we aim to *automatically identify segment boundaries that denote shifts of coverage* and, in this manner, extract temporal relationships for examination.

Our specific contributions are:

1) A novel application to studying Internet use in communities using the i-Neighbors system. The voluntary participation of i-Neighbors users enables us to gain significant insight into questions of engagement and deliberation.
2) Qualitative as well as quantitative summaries of distinctions observed between advantaged and disadvantaged communities. These results lead to an understanding of how engagement and deliberation practices relate to access and uses of new communication technologies.
3) A time series segmentation algorithm where segment boundaries detect significant shifts of topic coverage. To this purpose, we embed a topic modeling algorithm inside a segmentation algorithm and optimize for segment boundaries that reflect significant shifts of topic content.

II. Internet use in communities

This study builds on prior research that explores the relationship between Internet use and local engagement [20], [21], [22], [23], [24], [25]. In particular, we focus on the uneven impacts that Internet use may have on participatory democracy and informal deliberation for communities with a concentration of poverty.

A number of studies have demonstrated that the availability of a relatively simple neighborhood website and discussion forum can increase local tie formation, informal deliberation, and civil and civic behaviors [21], [20], [25]. For example, a longitudinal study of how local social networks changed as a result of a neighborhood email list found that the average person gained over four new local social ties for each year that they used the intervention [21]. Moreover, the type of discussion that was common in these forums was found to promote collective action and civic engagement [20], [21]. A recent, large, random survey of American adults found that of those who use an online neighborhood discussion forum, 60% know all or most of their neighbors, 79% talk with neighbors in person at least once a month, and 70% had listened to a neighbor‘s problems in the previous six months. This compared to the average American, 40% of whom knew their neighbors, 61% talked in-person, and 40% listened to a neighbor‘s problems [26].

The i-Neighbors system (Fig. 1) was created as part of a university research project first run from the Massachusetts Institute of Technology and later from the University of Pennsylvania, that has been operational since 2004 [25]. The site allows anyone in the United States or Canada to join and create a virtual community that matches their geographic neighborhood. Users who join the website agree to a Terms of Use, as approved by the Institutional Review Board (IRB). Through the Term of Use, users are informed that participation is voluntary and that logs of user activity would be recorded and analyzed. The i-Neighbors project was designed as a naturalistic experiment; there was no attempt to provide training or to encourage any individual user or community to participate. The website offers the following services:

- Discussion forum / email list: each neighborhood has a discussion forum that allows users to contribute and comment by email.
- Directory: a list of all group members and their profile information.
- Events calendar: a group calendar.
- Photo gallery: a group photo gallery.
- Reviews: user contributed reviews of local companies and services.
- Polls: surveys administered to other group members.

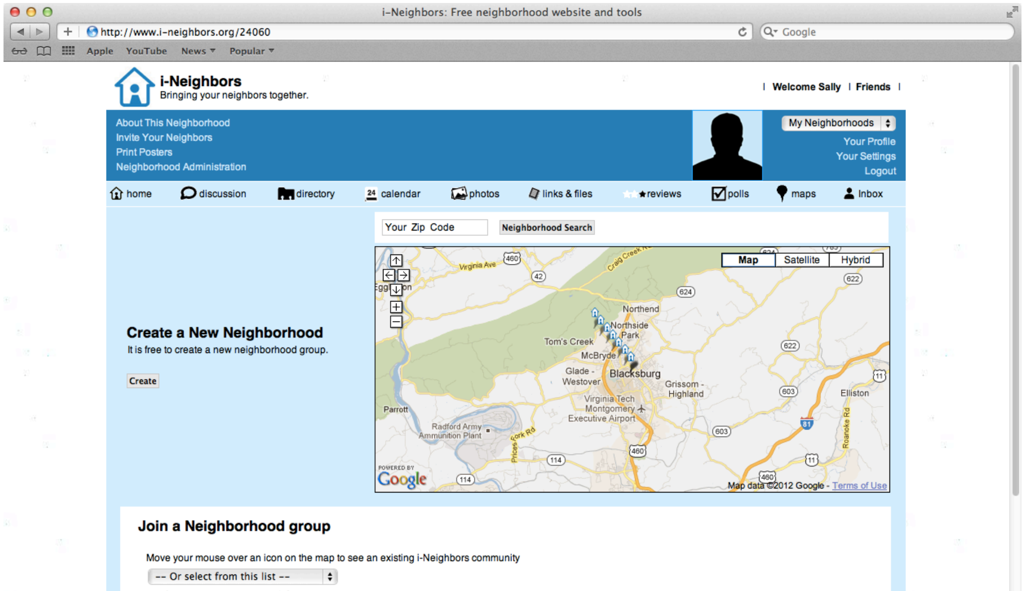

Fig. 1. i-Neighbors: Social networking service connecting residents of geographic neighborhoods.

- Documents: storage for shared documents and links.

As of 2012, the i-Neighbors website has attracted over 110,000 users who have registered over 15,000 neighborhoods. The size of each group and the number of active groups varies from month to month. In a typical month, over 1,000 neighborhoods are active and over 7,000 unique messages are collectively contributed to neighborhood discussion forums, which in turn are viewed over 1 million times. This analysis focuses on the adoption pattern of the most active i-Neighbors communities, based on measures of the concentration of poverty, and the content of messages contributed to their respective discussion forums.

III. Temporal Topic Modeling

As mentioned in the introduction, classical LDA does not model temporal evolution and there are many variants of LDA that do capture trends over time. Nevertheless, our needs here go beyond these systems, since we aim to *automatically identify segment boundaries that denote shifts of coverage* and, in this manner, extract temporal relationships for examination. Temporal topic modeling began to grab the attention of machine learning researchers around the beginning of 2006, with a fair amount of work being expended in this space. Although our specific problem is different, it is helpful to survey this thread of work with a view toward understanding commonalities of purposes.

Many existing temporal topic modeling algorithms modify the topic modeling algorithm itself to enable tracking topics over time. For instance, Blei and Lafferty [17] extended the classical state space models to identify a statistical model of topic evolution. In particular, they use state space models on the natural parameters of the underlying topic multinomial and on the natural parameters for the logistic normal distributions used for specifying the document-specific topic proportions. They also developed techniques for approximating the posterior inference for detecting topic evolution in a document collection.

Wang and MacCallum [27] have proposed a non-Markov model for detecting topic evolution over time. They assume that topics are associated with a continuous distribution over timestamps and that the mixture distribution over topics that represent documents is influenced by both word co-occurrence relationships and the document timestamp. In their model, thus, topics generate both observed timestamps and words. Iwata and Yamada [19] have proposed a topic model that enables sequential analysis of the dynamics of multiple time scale topics. In their proposed model, topic distributions over words are assumed to be generated based on estimates of multiple timescale word distributions of the previous time period. Wang and Blei [28] have recently proposed a model that replaces the discrete state space that was originally proposed in Blei and Lafferty [17] but with a Brownian motion law [29] to model topic evolution. They assume that topics are divided into sequential groups so that topics in each slice are assumed to evolve from the previous slice. Some recent papers have targeted the goal of modeling multiple information sources along with capturing topic evolution. Zhang et al. [18] have proposed an evolutionary hierarchal Dirichlet process (EvoHDP) model which extends the hierarchical Dirichlet process (HDP) to take time into account [30]. Inference of EvoHDPs is conducted through a cascade Gibbs sampling strategy. Hong, Dom, Gurumurthy, and Tsioutsiouliklis [31] have also addressed multiple streams and the temporal dynamics of topics detected from these streams. They tackle the multiple stream problem by allowing each text stream to have both local topics and shared topics. Each topic is associated with a function that characterize the topic popularity over time and this function is time-dependent. Some of the previous mentioned research along with others have focused on detecting topics from steaming data [32], [33], [34]; this is an important issue, but not the focus of this paper.

IV. Dynamic Temporal Segmentation Algorithm over Topic Models

To contrast our approach with the above, our goal is to not simply track temporal evolution of topics but to identify segments that denote significant shifts of content (distributions). In turn, this will help to detect differences in deliberation and common interests between advantaged and disadvantaged neighborhoods. This requires us to capture similarities and distinctions between neighborhoods based on: the amount of time neighborhoods with different poverty levels spent discussing the same topics, average similarity in topics discussed between neighborhoods with different poverty levels, and average similarity in topics discussed between neighborhoods with the same poverty levels.

To characterize discussions within neighborhoods we developed a novel integration of segmentation algorithms with topic modeling algorithms. We aim to identify segmentations such that segment boundaries indicate qualitative changes in topic distributions. (Note that this goal is different from classical temporal topic modeling because we seek to identify 'break points' where significant shifts of topic are occurring.) Every neighborhood in our analysis is characterized in this manner and the resulting segmentations are then clustered with a view toward identifying enrichments that hold (or do not) at different poverty levels.

A. Segmentations driven by topic dynamics

Our first task was to segment the time course such that segment boundaries indicated important periods of temporal evolution and re-organization. We operationalized this notion in the following manner: if topic modeling were to be conducted on either side of a segment boundary, the discovered topics should be qualitatively different from each other (and, thus, the boundary captured a significant shift in word distributions). Building upon our prior work, to realize these objectives we embedded an collapsed Gibbs sampling based LDA used for discovering topics into a time series segmentation algorithm [35], [36] .

There are many algorithms available for topic modeling of text, e.g., pLSI [37] and LDA [38]. Typically in a topic model, a topic is considered a distribution over words and a document is in turn modeled as a distribution over topics. LDA, in particular, assumes that documents are generated in two stages: (i) specify a distribution over topics, (ii) to generate words for a given document, sample a topic and then sample words from the chosen topic's distribution of words; repeat as necessary.

Given *N*, the number of words, *M*, the number of documents, θ denoting the topic mixture, β and α being Dirichlet parameters, z being the topic assignments, and w denoting the words, the generative process for LDA can be expressed mathematically as a joint probability distribution over the observed documents, topic structure, per-topic-document topic distribution, and per-document per-word topic assignments:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \quad (1)$$

The α is used to generate the topic mixture θ and β is used to generate the word probabilities. In our implementation the α and β values were set to constants ($\alpha = 0.01$) and ($\beta = 0.01$) (defaults used in the Stanford Topic Modeling Toolbox [39]).

The goal of LDA inference is to uncover the underlying topic structure parameters using only the observed variables w. Once the hyper-parameters α and β are inferred, reconstructing the generative process for a new set of documents involves a simple application of Bayes rule:

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)} \quad (2)$$

The next step is to take time-indexed documents as input and identify segment boundaries automatically: Given data indexed over a time series $T = \{t_1, t_2, ..., t_t\}$, the segmentation problem we are trying to tackle is to express T as a sequence of segments or windows: $(S_{t_1}^{t_a}, S_{t_{a+1}}^{t_b}, \ldots, S_{t_k}^{t_l})$ where each of the windows $S_{t_s}^{t_e}, t_s \leq t_e$, denotes a contiguous sequence of time points with t_s as the beginning time point and t_e as the ending time point.

Our operating assumption is that if we were to apply LDA separately on each side of a segment boundary, we should encounter independent topics. The notion of independence is captured here using a contingency table formulation as shown in Fig. 2. This figure illustrates activity around a putative segment boundary. Either side of the boundary demonstrates topic models specific to that segment (recall these are distributions over words/terms). Edges are drawn to illustrate term movement/reorganization across the boundary. The goal is to maximize such reorganization so that terms dynamically restructure into other topics or terms appear or disappear over time.

We compare segments by comparing the underlying topic distributions and quantifying common terms and their probabilities. There are many ways to accomplish this objective, e.g., by comparing overlaps between top-k terms, or by inner product measures over probability distributions. In either approach, the resulting overlap between distributions can be captured in the form of a contingency table (as shown in Fig. 2). If the contingency table entries are near uniform, it means that the two topic models are maximally independent and that we have arrived at a good segmentation boundary.

There are two aspects that have to be formulated now. First, how do we quantify the uniformity of the contingency table? Second how do we identify the segment boundaries automatically?

We begin by describing how we evaluate two adjacent windows, assuming the segmentation boundary is given. Then we outline how to automatically identify segmentation boundaries. Let the two adjacent windows be $S_{t_a}^{t_b}$ and $S_{t_{b+1}}^{t_c}$. As mentioned

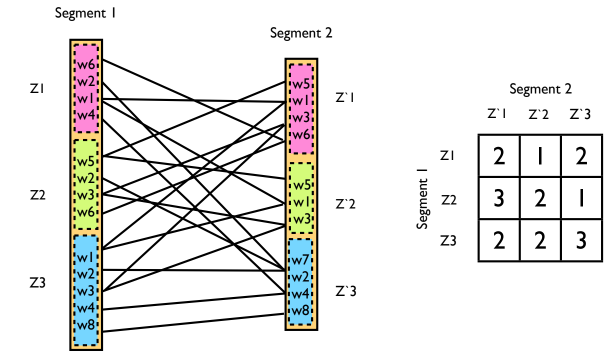

Fig. 2. Contingency table used to evaluate independence of topic distributions across a putative segment boundary.

before, we setup a contingency table of size $r \times c$ where rows r denote topics in one window and columns c denote topics in the other window. Entry n_{ij} in cell (i, j) of the table represents the overlap of terms between topic *i* of $S_{t_a}^{t_b}$ and topic *j* of $S_{t_{b+1}}^{t_c}$.

We first calculate two auxiliary quantities:

- Column-wise sums (representing the sizes of clusters in $S_{t_a}^{t_b}$) across each row: $n_{i.} = \sum_j n_{ij}$
- Row-wise sums (representing the sizes of clusters in $S_{t_{b+1}}^{t_c}$) down each column: $n_{.j} = \sum_i n_{ij}$

From these two calculations we define (r+c) probability distributions, one for each row and one for each column.

$$p(R_i = i) = \frac{n_{ij}}{n_{i.}}, (1 \leq j \leq c) \tag{3}$$

$$p(C_j = i) = \frac{n_{ij}}{n_{.j}}, (1 \leq i \leq r) \tag{4}$$

Now we formulate the objective function F to capture the deviation of these row-wise and column-wise distributions w.r.t. the uniform distribution:

$$F = \frac{1}{r} \sum_{i=1}^{r} D_{KL}(R_i \| U(\frac{1}{c})) + \frac{1}{c} \sum_{j=1}^{c} D_{KL}(C_j \| U(\frac{1}{r})) \tag{5}$$

where

$$D_{KL}(R_i \| U(\frac{1}{c})) = \sum_i p(R_i) \log \frac{p(R_i)}{p(U(\frac{1}{c}))} \tag{6}$$

$$D_{KL}(C_j \| U(\frac{1}{r})) = \sum_j p(C_j) \log \frac{p(C_j)}{p(U(\frac{1}{r}))} \tag{7}$$

Here D_{KL} denotes the KL-divergence. Note that the KL-divergence is not a true distance measure; for instance, it is not symmetric. The goal of the optimization is to minimize F, in which case the distributions observed in the contingency table are as close to a uniform distribution as possible, in turn implying that the topics are maximally dissimilar.

The algorithm automatically identifies segmentation boundaries by first checking the start and end dates of the available data. It uses then two sliding windows that move across the data which are evaluated by computing the previously described objective function F. The sizes of the two sliding windows are bounded by a minimum and maximum window sizes specified as input parameters for the algorithm. The algorithm evaluates all permutations of the two sliding window sizes and adds a segmentation point when the objective function is optimized or when we reach the maximum window size for both windows.

B. Characterizing Neighborhoods

We use the above segmentation algorithm to track discussions across each individual neighborhood; the next step is to compare such segmentations across neighborhoods.

Recall that since LDA topics are characterized in terms of distributions over terms ($p(w|z_n)$) and that such distributions are weighted to yield the joint distribution:

$$p(w, z_n) = p(z_n).p(w|z_n) \tag{8}$$

These distributions (one for each segment of each neighborhood) must now be compared with an aim toward identifying commonalities and discrepancies. However, before we capture distinctions between such distributions, we must ensure that the underlying distributions are expressed over the same vocabulary (terms). To this end, we use the superset of terms from both distributions as the sample space over which two segments induce their respective distributions.

Most clustering algorithms require a symmetric measure of association and we employ the Jensen-Shannon Divergence (JSD):

$$JSD(P \| Q) = \frac{1}{2} D_{KL}(P \| M) + \frac{1}{2} D_{KL}(Q \| M) \tag{9}$$

where

$$M = \frac{1}{2}(P + Q) \tag{10}$$

Note that the Jensen-Shannon divergence is just a symmetrized version of the KL-divergence. The dissimilarity matrix constructed in this manner can be used as input to any clustering algorithm, e.g. an agglomerative clustering with single-linkage criterion is used here.

V. Methods

To test our hypothesis, that social media can afford democratic engagement in areas of concentrated poverty, we focus our analysis on where the i-Neighbors intervention has been a success. By focusing on the 20 most active i-Neighbors groups, we identify local areas that have successfully adopted social media for civic and civil engagement (as shown in Fig. 3). Traditionally, we would expect to find very few examples of engagement in areas where poverty rates are high - nearly all successful i-Neighbors groups should be in areas where there is little concentration of inequality. However, our hypothesis runs counter to this traditional expectation, we expect social media to afford successful democratic engagement in areas where poverty rates are high.

We ranked neighborhoods based on the number of unique comments that members posted to their neighborhood‘s discussion forum over a one year period that started on October 1, 2010. For each neighborhood group, we identified the poverty rate, as defined by the US Census [40], based on Census tract data collected as part of the 2009 American Community Survey (US Census Bureau). In Fig. 4 the same neighborhoods shown in Fig. 3 were rearranged based on poverty level. While recognizing that the selection of any absolute threshold will have its shortcomings, consistent with previous research, we used a 20% poverty rate as an indicator of an area of high-poverty [41].

The percentage of families below the poverty level in geographic areas represented by the 20 most active i-Neighbors groups ranges from a low of 3.2% to a high of 47.6%. 40% of the most active neighborhoods are in areas of concentrated poverty. Given that 15% of Americans live below the poverty level [41], that 40% of the most active i-Neighbors groups are in areas where more than 20% of families are in poverty indicates adoption by high poverty neighborhoods at a higher rate than would be expected at random. This finding supports our hypothesis that social media afford opportunities for democratic engagement in areas of concentrated poverty, at a rate that is as high (or higher) than more advantaged areas.

To test our hypothesis that informal deliberation in areas of high poverty would be similar to deliberation that takes place in areas where poverty is low, we modeled how long neighborhoods with different poverty levels spent discussing topics (Section VI.A), the average similarity in topics discussed between neighborhoods with different poverty levels, and the average similarity in topics discussed between neighborhoods of similar poverty levels (Section VI.B). To facilitate interpretation, we limited the scope of our analysis to the three most active i-Neighbors groups above our 20% poverty level threshold, and the three most active below the threshold. While we recognize that there are a number of potential sampling approaches, including sampling groups from similar or diverse geographic areas, we chose to maximize the available data for topic modeling. However, our approach also served to provide a sample that was geographically diverse, with the six groups used for our topic analysis representing six different U.S. States as shown in Table I .

Our goal is to study two basic questions:

- What lengths of time neighborhoods with different poverty levels spend discussing topics? (Section VI.A)
- What is the average similarity in topics discussed between neighborhoods with different poverty levels, and the average similarity in topics discussed between neighborhoods with similar poverty levels? (Section VI.B)

VI. Findings

We applied our temporal segmentation algorithm on the six selected neighborhoods. The output of the algorithm is a set of segments from each neighborhood, a dissimilarity matrix, and a dendrogram depicting the clustering of all segments across neighborhoods. Some segments were examined manually, by checking the original text to validate the segmentation output. A partial segmentation output is shown in Fig. 6 for a disadvantaged neighborhood and in Fig. 5 and Fig. 7 for a more advantaged neighborhood.

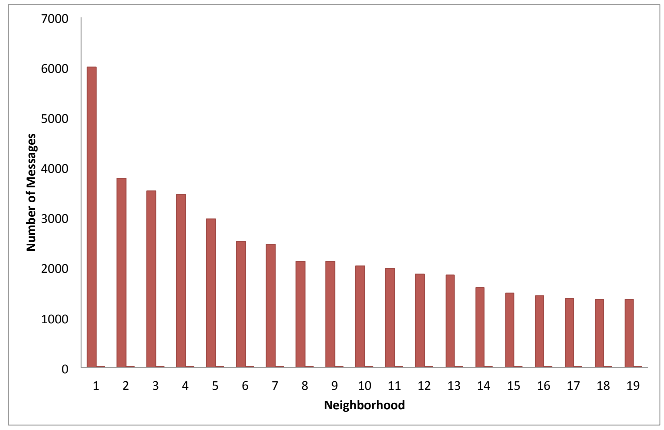

Fig. 3. Distribution of messages across neighborhoods.

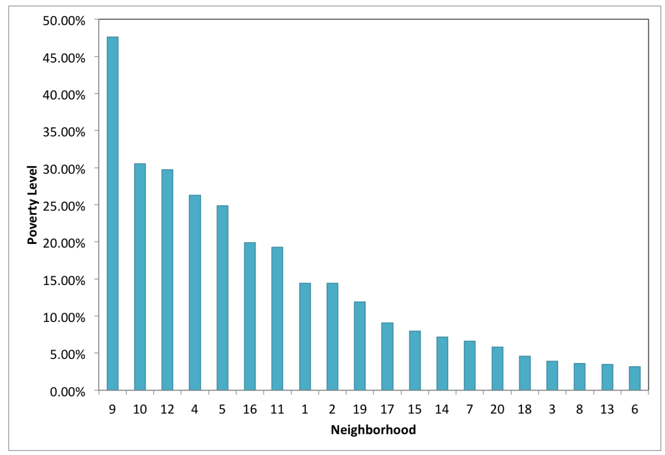

Fig. 4. Distribution of poverty levels in neighborhoods.

A. Characterizing Segment Durations

Fig. 8 depicts the segmentation outputs for the six disadvantaged and advantaged neighborhoods for the one year period in which messages were exchanged within the communities. The segmentation algorithm was applied on each neighborhood separately to identify shifts in topics. Segments identified from each neighborhood are aligned so that vertical ordinates denote the same time point globally. The dashed vertical lines in each segmentation denote the algorithm-picked boundaries. There is not a significant difference in segment durations across the two classes of neighborhoods. The average length of segments from advantaged neighborhoods is 3.24 months, whereas the average length of segments from disadvantaged neighborhoods is 3.38 months. (Note that the segment features a collection of topics during its time, but this does not mean that all these topics were discussed during the entire duration of the segment.)

TABLE I
THE SIX NEIGHBORHOODS STUDIED IN OUR EXPERIMENTS.

Neighborhood ID	Number of Members	Number of messages	State	Poverty
High1	440	2122	Ohio	47.60%
High2	334	3466	New York	26.30%
High3	539	2969	Maryland	24.90%
Low3	378	2472	Texas	6.60%
Low2	324	3534	Georgia	3.90%
Low1	371	2523	North Carolina	3.20%

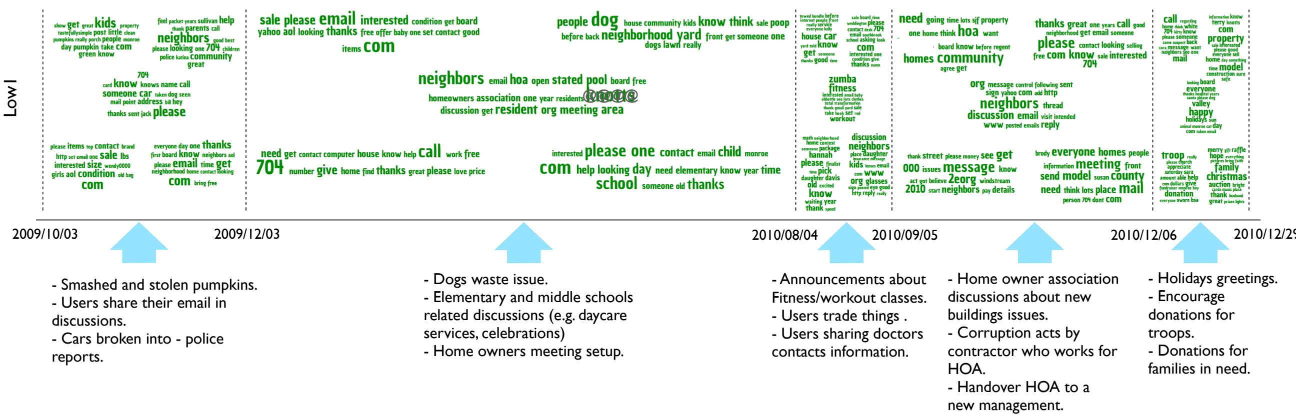

Fig. 5. Partial segmentation output from a low-poverty neighborhood.

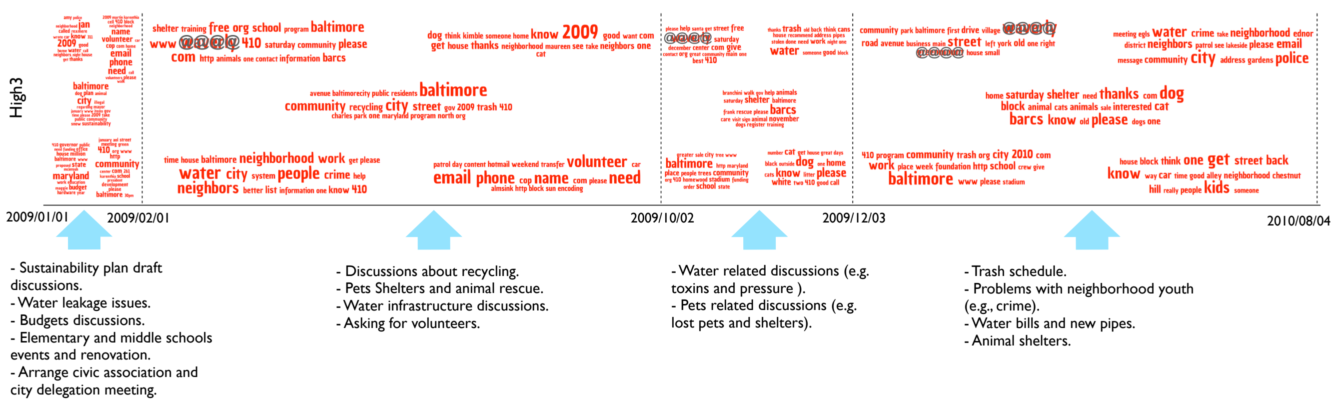

Fig. 6. Partial segmentation output from a high-poverty neighborhood.

B. Characterizing Topical Content of Segments

We employed our inferred topic models to construct the dissimilarity matrix across neighborhood segments using the approach described earlier. Topics ranged in similarity from 0 to 4.43, where zero means that the two segments are identical.

If discussion topics within disadvantaged neighborhoods were substantively different from topics within neighborhoods that have lower poverty levels, the divergence coefficient would be significantly higher between advantaged and disadvantaged neighborhoods than it is within neighborhoods that are similar in poverty. That is, we would expect topics within neighborhoods of similar poverty level to be more similar to each other than they are with neighborhoods that are substantively different in poverty.

Across neighborhoods, dissimilarity in segments ranges from 0 (identical) to 4.43, the mean difference is 2.19 ($SD = 1.09$). The mean divergence coefficient between all discussion topic pairs within communities that are low in poverty is 2.18 ($SD = 1.09$), ranging from 0.11 to 4.42. The average divergence *between* all neighborhoods low in poverty is not significantly different from the average divergence of topics *within* neighborhoods low in poverty ($M = 2.11$,

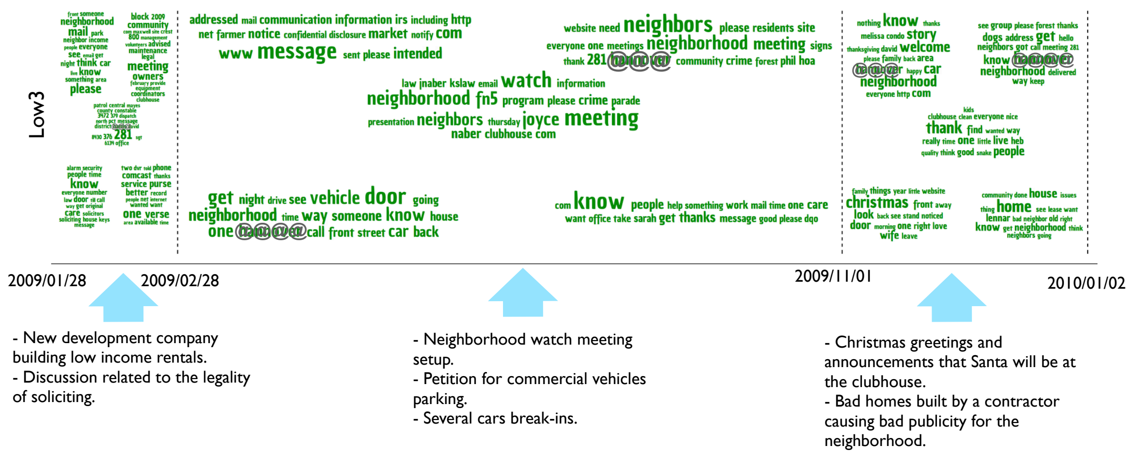

Fig. 7. Partial segmentation output from a low-poverty neighborhood.

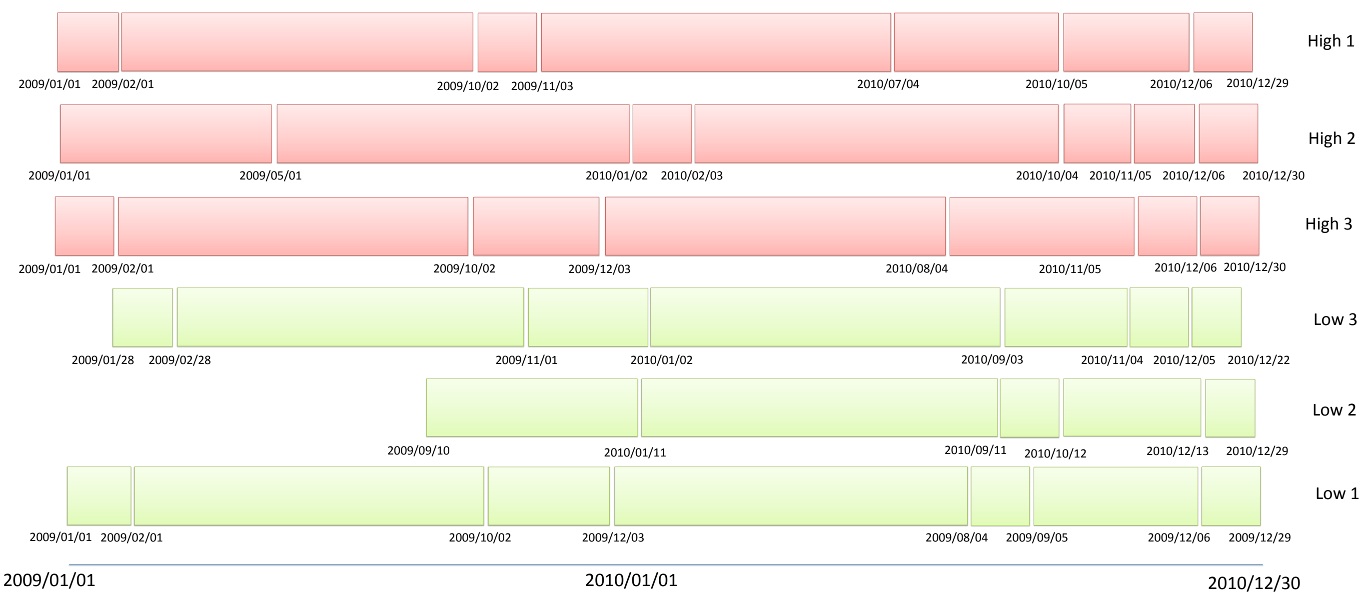

Fig. 8. Durations of segments in advantaged (low poverty) and disadvantaged (high poverty) neighborhoods.

SD=1.01; one-way ANOVA $> .05$). Topics discussed within low poverty neighborhoods are similar across all low poverty neighborhoods.

The average divergence coefficient between all topic pairs across all high poverty areas ranges from 0.09 to 4.20 with a mean of 2.26 ($SD = 1.09$). Looking *within* high poverty neighborhoods, the mean divergence is 2.35 ($SD = 1.07$), which is not significantly different from the divergence *between* topics in similar high poverty areas ($M = 2.21$, $SD = 1.10$; one-way ANOVA > 0.05). The variation in topics discussed within high poverty neighborhoods is consistent across high poverty neighborhoods.

Comparing discussion topics in high and low poverty areas, divergence ranges from 0.20 to 4.43 with a mean divergence of 2.16 ($SD = 1.09$). There was no significant difference between divergence *within* neighborhoods of similar poverty level in comparison to divergence *between* neighborhoods of contrasting poverty (one-way ANOVA > 0.05). Consistent with our hypothesis, the variation in topics discussed within advantaged and disadvantaged areas is not statistically different than the variation in topics between areas of high and low poverty. The range and nature of topics is the same in high poverty areas as was found in more advantaged areas.

A flat clustering of segments reveals congruences as well as outliers. Fig. 9 depicts some segments that were clustered together and the topics that contributed to their clustering. Other outliers segments are also shown in the figure. Non-outliers reveal common discussions about topics.

For example, in Neighborhood 7 [2009-03-01 - 2009-11-01] and Neighborhood 4 [2009-01-01 - 2009-05-01], there were messages discussing the setup of a neighborhood watch meeting and messages discussing a petition. The petition was

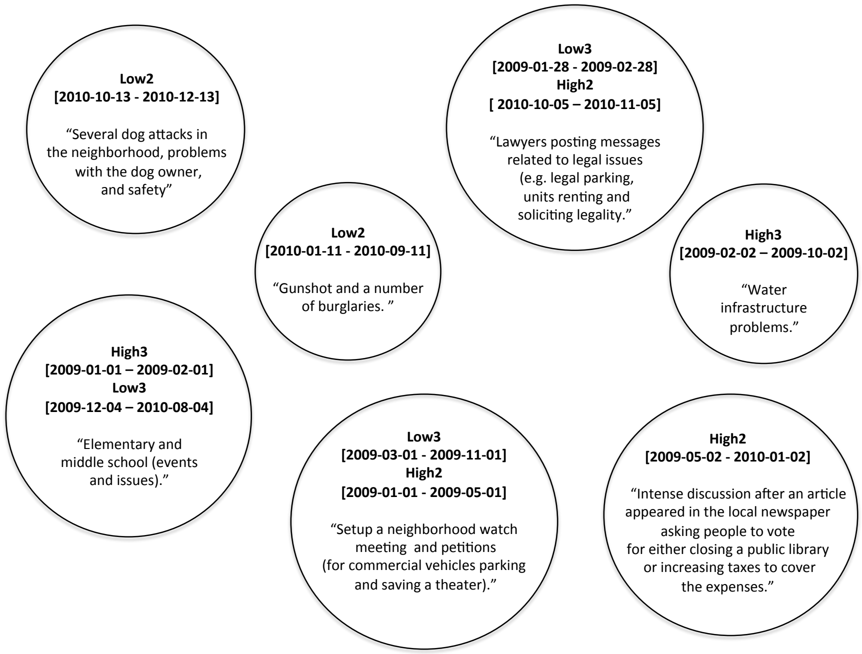

Fig. 9. Example clusters of discovered segments across neighborhoods.

for commercial vehicles parking in Neighborhood 7 and in Neighborhood 4 it was to save a theater. In Neighborhood 5 [2009-01-01 - 2009-02-01] and Neighborhood 6 [2009-12-04 - 2010-08-04], there were many messages about elementary and middle school events and issues. On the other hand, outliers reveals discussions about an unusual topic. For example, in Neighborhood 3 [2010-01-11 - 2010-09-11], we found a lot of messages discussing a gunshot and a number of burglaries. In this segment, a lot of messages discuss how to buy a gun or a dog. Another example is Neighborhood 4 [2009-05-02 - 2010-01-02], which had an intense discussion after an article appeared in the local newspaper asking people to vote for either closing a public library or increasing taxes to cover the expenses. The last example is Neighborhood 3 [2010-10-13 - 2010-12-13], which had many messages discussing several dog attacks in the neighborhood, problems with the dog owner, and safety.

VII. Conclusion and Future Work

In this paper we address the divide in democratic engagement that exists between advantaged and disadvantaged communities. We look for evidence that the gap between high and low poverty communities, in democratic participation and deliberation, is affected by the use of a social media intervention. Specifically, we have argued that new communication technologies afford civic and civil behaviors and informal deliberation in high poverty communities, similar to what is experienced in communities that are low in poverty. Our approach compares the adoption of a new technology across neighborhoods of high and low poverty. We use a unique algorithm to:

- Detect differences in deliberations activity between neighborhoods with different poverty levels.
- Detect whether there are more or less common discussion topics between communities with different poverty levels.

We did not find significant differences between high and low poverty neighborhoods in terms of either the lengthy of discussion periods or the overall topics of discussion. In addition, we found that the rate of adoption of a communication tool for participatory democracy was much higher than would be expected based on established theories pertaining to the digital divide and concentrated inequality. This is not the usual finding in studies of the digital divide, where lower socioeconomic status populations typically have fewer opportunities to participate in public deliberation.

In the past structural constraints internal to disadvantaged communities limited opportunities for deliberation and democratic participation. Social technologies may make communication possible where it was not before. One possible explanation, as to why social media may be such an important tool for engagement among this population, may relate to the way these technologies bring people together. Previous findings, that use of the Internet as an information tool has a modest positive relationship to engagement for those who are already likely to be engaged [10], [11], [13], do not extend to the truly disadvantaged. However, when the Internet is used as a social tool, a means to communication between people who are “locally” embedded in existing social structures (even if those structures are loosely connected) it affords social cohesion, discussion, and engagement. Technologies that facilitate communication among a population that shares geography, or possibly other sources of affiliation, enables contact that may previously have been desired, but was constrained by physical and structural barriers. It may not be surprising that, when barriers to contact are reduced, we find that residents of high poverty areas are as motivated to participate and deliberate about local issues as people of other communities. If these findings are generalizable, the policy implications are significant. Insuring equal access to social media, across socioeconomic divides, has the potential to reduce persistent inequalities in democratic engagement.

Our next steps are to extend our segmentation algorithm to capture not just topic differences but sentiment evolutions. This will enable us to measure differences in public perception and attitudes between advantaged and disadvantaged neighborhoods.

References

[1] J. Helliwell and R. Putnam, "The social context of well-being," in *philosophical transactions of the royal society of London*, vol. 359, no. 1449. Springer, 2004, pp. 1435–1446.

[2] M. X. D. Carpini and S. Keeter, *What americans know about politics and why it matters*. Yale University Press, 1996.

[3] R. Putnam, *Bowling alone*. Simon and Schuster., 2000.

[4] E. Klinenberg, "Heat wave." The university of Chicago press, 2002.

[5] J. S. Fishkin, *The voice of the people*. Yale university press, 1995.

[6] K. C. Walsh, "Talking about politics: informal groups and social identity in american life." The university of Chicago press books, 1992.

[7] J. Dewey, *The public and its problems*, 1st ed. Swallow press, 1927.

[8] E. M. Uslaner and M. Brown, "Inequality, trust, and civic engagement," 2005, pp. 868–894.

[9] R. J. Sampson, "Great american city: chicago and the enduring neighborhood effect." The university of chicago press, 2011.

[10] D. V. Shah, J. Cho, W. P. J. R. Eveland, and N. Kwak, "Information and expression in a digital age," in *communication research*, vol. 32, no. 5, 2005, pp. 531–565.

[11] S. Boulianne, "Does internet use affect engagement? a meta-analysis of research," in *political communication*, vol. 26, no. 2, 2009, pp. 193–211.

[12] K. N. Hampton, L. S. Goulet, L. Rainie, and K. S. Purcell, "Social networking sites and our lives: how people's trust, personal relationships, and civic and political involvement are connected to their use of social networking sites and other technologies." in *Public sociology: research, action, and change.* Pew research center, 2011.

[13] E. Hargittai and A. Shaw, "The internet, young adults and political engagement around the 2008 presidential election," Presented at the berkman center for internet and society at Harvard university, 2011.

[14] R. Merton, "The matthew effect in science," in *Science*, vol. 159, no. 3810, 1968, pp. 56–63.

[15] A. Kavanaugh, *The arc of social computing: interaction in web versus physical communities (to appear)*. springer, 2013.

[16] D. M. Blei, "Introduction to probabilistic topic models," 2011.

[17] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *The 23rd international conference on machine learning, ICML*, 2006, pp. 113–120.

[18] J. Zhang, Y. Song, C. Zhang, and S. Liu, "Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora," in *proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining*, 2010, pp. 1079–1088.

[19] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda, "Online multiscale dynamic topic models," in *proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining*, 2010, pp. 663–672.

[20] K. N. Hampton and B. Wellman, "Neighboring in netville: how the internet supports community and social capital in a wired suburb." in *city and community*, vol. 2, no. 4, 2003, pp. 277–311.

[21] K. N. Hampton, "Neighborhoods in the network society: the e-neighbors study," in *information, communication and society*, vol. 10, no. 5, 2007, pp. 714–748.

[22] A. Kavanaugh, A. Cohill, and S. Patterson, "The use and impact of the blacksburg electronic village, in community networks: lessons from blacksburg, virginia." Artech house, 2000.

[23] A. Kavanaugh, T. Zin, M. Rosson, J. Carroll, J. Schmitz, and B. Kim, "Local groups online: political learning and participation." in *computer supported cooperative work*, vol. 16, no. 4-5, 2007, pp. 375–395.

[24] A. Kavanaugh, B. Kim, J. Schmitz, and M. Pérez-Quiñones, "Net gains in political participation: secondary effects of the internet on community." in *information, communication and society*, vol. 11, no. 7, 2008, pp. 933–963.

[25] K. N. Hampton, "Internet use and the concentration of disadvantage: globalization and the urban underclass," in *american behavioral scientist*, vol. 53, no. 8. SAGE publications, 2010, pp. 1111–1132.

[26] K. N. Hampton, L. Sessions, E. J. Her, , and L. Rainie, "Social isolation and new technology: how the internet and mobile phones impact americans social networks," in *Pew research center*, 2009.

[27] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, 2006, pp. 424–433.

[28] C. Wang, D. Blei, and D. Heckerman, "Continuous time dynamic topic models," in *Proceedings of the 24th conference in uncertainty in artificial intelligence (UAI)*, 2008.

[29] G. Lawler, "Introduction to stochastic processes." Chapman & Hall, 1995.

[30] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," in *Journal of the american statistical association*, vol. 101, 2004.

[31] L. Hong, B. Dom, S. Gurumurthy, and K. Tsioutsiouliklis, "A time-dependent topic model for multiple text streams," in *proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*, 2011, pp. 832–840.

[32] A. Gohr, A. Hinneburg, R. Schult, and M. Spiliopoulou, "Topic evolution in a stream of documents." in *SDM'09*, 2009, pp. 859–872.

[33] M. Hoffman, D. M. Blei, and F. Bach, "Online learning for latent dirichlet allocation," in *advances in neural information processing systems 23*, 2010, pp. 856–864.

[34] L. AlSumait, D. Barbará, and C. Domeniconi, "On-line lda: adaptive topic models for mining text streams with applications to topic detection and tracking," in *proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 3–12.

[35] S. Tadepalli, N. Ramakrishnan, L. T. Watson, B. Mishra, and R. F. Helm, "Simultaneously segmenting multiple gene expression time courses by analyzing cluster dynamics," in *jornal of bioinformatics and computational biology*, vol. 7, no. 2, 2009, pp. 339–356.

[36] M. S. Hossain, S. Tadepalli, L. T. Watson, I. Davidson, R. F. Helm, and N. Ramakrishnan, "Unifying dependent clustering and disparate clustering for non-homogeneous data," in *proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining*, 2010, pp. 593–602.

[37] T. Hofmann, "Probabilistic latent semantic analysis," in *in proceeding of uncertainty in artificial intelligence, UAI'99*, 1999, pp. 289–296.

[38] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," in *journal of machine learning research*, vol. 3, 2003, pp. 993–1022.

[39] "Stanford topic modeling toolbox," http://nlp.stanford.edu/software/tmt/tmt-0.4/, July 2012.

[40] "Poverty," http://www.census.gov/hhes/www/poverty/methods/definitions.html, July 2012.

[41] E. Kneebone and A. Nadeau, C.and Berube, "The re-emergence of concentrated poverty: metropolitan trends in the 2000s," ser. Metropolitan opportunity series. Brookings, 2011.