Social Activity Hubs

Estimating User Specific Contextual Factors from Social Media Data

Ted Hsuan Yun Chen
Department of Political Science
Pennsylvania State University
State College, Pennsylvania
thc126@psu.edu

Paul Zachary
Institute for Quantitative Theory and Methods
Emory University
Atlanta, Georgia
paul.zachary@emory.edu

Christopher J. Fariss
Department of Political Science
University of Michigan
Ann Arbor, Michigan
cjfariss@umich.edu

ABSTRACT

Context influences sociopolitical attitudes and behaviors, making the estimation of individuals' contexts an important methodological problem for the social sciences. We add to this body of work by presenting a method to estimate an individual's spatial contexts, specifically the set of geospatial areas an individual is most active in. Our approach, which utilizes the Dirichlet process mixture model, departs most significantly from more traditional approaches to estimating relevant spatial locations in that it does not arbitrarily constrain the number of spatial contexts an individual can have. This modeling approach reflects our recognition that an individual's lived experiences is a combination of different contexts that overlap to varying degrees. This flexibility therefore yields a more valid measure of spatial contexts. To illustrate our method, including its performance relative to other measures, we apply our method to Twitter data generated by protesters who participated in the 2015 Freddie Gray protests in Baltimore, MD.

CCS CONCEPTS

• **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → *Machine learning*; *Machine learning approaches*;

KEYWORDS

Social networks, geographic profiling, text analysis

ACM Reference Format:
Ted Hsuan Yun Chen, Paul Zachary, and Christopher J. Fariss. 2017. Social Activity Hubs: Estimating User Specific Contextual Factors from Social Media Data. In *Proceedings of CSSSA's Annual Conference on Computational Social Science.* ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/10.1145/3145574.3145606

1 INTRODUCTION

Across a wide variety of social sciences, context has been repeatedly shown to be one of the most important determinants of human social behavior. Context matters. Individuals modulate their responses to stimuli according to their social network, the formality of the situation, and power and status differentials. While qualitative accounts of behavior have long emphasized the importance of context, recent empirical research provides causal evidence for the effect of context on behavior. Using a variety of sophisticated research designs, context has been shown to affect altruism [29]; anti-immigrant sentiment and discriminatory behavior [10–12]; and support for extremist politicians [17].

The relationship between context and behavior in the social sciences is not always easy to discern. Each individual's behavioral responses are conditioned by the social, economic, and political realities of their specific setting. Unfortunately, identification of individual-level contextual factors that are hypothesized to influence behavior are difficult to gather. Furthermore, it is often the case that the act of measurement itself changes the behavior of the subject under study. Measurement is typically an inherently social act. Survey participation requires the active and informed consent of the individual participant. In this context, social desirability bias is quite difficult to avoid. Thus, contextual factors at the individual level are difficult and costly to measure and often too complex to manipulate experimentally.

To overcome these challenges, many of the aforementioned studies rely upon some form of randomized intervention as a design feature, be it through the lab, in the field, or naturally occurring. In these studies, each subject's context is manipulated through various stimuli. For example, in one study participants encountered a homeless person, staged by the experimenters, as they completed a survey. This allowed the experimenters to measure the way in which visible poverty affects altruistic sentiment [29]. Another study engaged participants who self-identified as white/Caucasian in a game in which they were asked to play the role of a country's dictator. When these participants had been previously informed about the growth of the Hispanic population inside the United States, they were more likely to favor other white participants [1]. Research has also found that transphobia —prejudice against transgender people — is reduced when canvassers randomly encourage active-perspective-taking among survey respondents [6]. These studies collectively illustrate the powerful way in which situational factors can influence social behavior.

Randomized experiments convincingly establish causality and rule out alternate explanations. However, the external validity of such findings are not always easily established [13, 30, 31]. This problem is particularly acute in experiments that attempt to measure political and social context, which, by necessity, tend to adopt

CSSSA's Annual Conference on Computational Social Science, October 19–22, 2017, Santa Fe, NM, USA

ACM ISBN 978-1-4503-5269-7/17/10…$15.00
https://doi.org/10.1145/3145574.3145606

extremely powerful treatments in order to generate statistically detectable effects within limited sample sizes. In contrast, though observational studies do not typically have as strong internal validity, they effectively address the question of external validity. Observational studies measure subject specific-behaviors in real life situations, often without direct intervention on behalf of the experimenters, which may unintentionally alter the behavior under study [13]. For example, observational studies of political context have considered the way in which demographics affect "white flight" [9, 23], diversity and inter-ethnic violence [7], and voter turnout for minority candidates [3].

Most theories about behavioral variations across contexts operate at the individual level, but inference at this level of analysis is often constrained by two primary difficulties associated with data availability. First, sparsity in behavioral responses measured at the individual level often leads to the use of group-level outcomes as aggregated behavioral measures. The resulting inference may be subject to a serious ecological inference problem [22]. Second, subject to data constraints, researchers are often forced to rely on coarse measures of context. Depending on the particular mechanisms specified in the study, the use of coarse, non-localized contexts may be inappropriate as there might be high levels of within-context variation in the stimuli of interest. Imprecise measures of context therefore introduce high levels of measurement error, rendering the inferred contextual factors not meaningfully relevant to individuals' experiences. Scholars have adopted a variety of strategies to address this issue of ecological validity, such as using small units of analysis when available [7], or adopting different statistical methods. The increasingly widespread penetration of social media, and the resulting flood of individual-level data, presents a third option: measuring context through social media behavior [24, 25].

An increasingly rich and important literature uses social media to draw inferences about contextual effects across a variety of domains and behaviors, including job-seeking [14], voting behavior [5], collective action in autocracies [32, 33], and even politicians' communication with one another [2]. Commonly-available forms of social media data still pose restrictions on the types of contextual effects that researchers can study. Although the application program interface (API) for some companies and commercial vendors provides information about users' time zone or country, these are measures of context only in the broadest and most aggregate sense. Using these locations to infer context potentially creates the same type of ecological inference issue as described above. Again, it is crucial to address this issue, as we know that an individual's context matters for understanding how she conditions her behaviors.

To address this inferential problem, there is a burgeoning field of research that develops new methods for analyzing what is referred to as "volunteered geographic information" [21]. Grace et al. [18], for example, uses users' network ties to local organizations to infer residential location. Others, such as Hasan et al. [19], employ semantic analysis to extract geospatial information embedded in social media text. This field of research highlights the importance of continually developing means to infer context from social media data.

In this paper, we contribute to this developing line of research and present a method to estimate Twitter users' "social activity hubs" (SAHs), or the geospatial areas where users spend their time. Our approach, which builds on Rossmo [28] and Verity et al. [35], departs most significantly from earlier approaches in that it does not arbitrarily constrain the number of clusters an individual's overall movement profile can contain [35]. This modeling approach reflects our recognition that an individual's lived experiences is a combination of different contexts that overlap to varying degrees, and that data-driven methods of inferring how many contexts are relevant to each individual is superior to relying on assumptions when we lack strong *a priori* beliefs. We estimate these SAHs using an algorithm that selects between geoprofiling models of varying sophistication that are conditional on information availability. For users with enough information, we utilize posterior quantities from a Dirichlet process mixture model to compute SAHs.

As initial evidence the utility of SAHs as a measure of political context, we run our algorithm on a sample of geotagged tweets made by Twitter users who participated in the Freddie Gray protests in Baltimore, MD in April 2015. Estimated SAHs are plotted on a Baltimore City map in Figure 1. As is evident, our model yield SAHs that cluster in areas with high daytime populations, such downtown Baltimore, and the high population area surrounding Johns Hopkins University. This demonstrates that our SAHs are a useful tool to generate disaggregated, individual-level estimates of social media users' economic, social, or political contexts. In the remainder of this paper, we describe the SAH model, and further illustrate its applications using Twitter data from Baltimore during the Freddie Gray protests.

2 RESEARCH DESIGN

2.1 Background on the Social and Political Context during the 2015 Baltimore Protests

Freddie Gray was arrested on April 12, 2105 by Baltimore Police Department (BPD) officers for possession of what officers believed at the time was an illegal switchblade.[1] For reasons that are under dispute, Gray fell into a coma while being transported in a police van subsequent to his arrest. He never recovered from his injuries and died in a trauma center on April 19, 2015. Starting on April 18, 2015, protesters began gathering in front of Baltimore's Western district police department to denounce Gray's alleged mistreatment and BPD brutality. These protests grew steadily in size as media attention to Gray's case increased throughout the week. The protests continued to gain momentum and eventually reached several thousand people. The Maryland National Guard, responding to a declared state of emergency, was brought in to restore order to the city. A mandatory curfew was declared within Baltimore city limits from April 28 to May 3. As Chen et al. [8] argue that contact with police affects behavior, our empirical application is focused on these protests.

2.2 Sampling and Data Collection

Prior to beginning our analysis, we first purchased all of the geotagged tweets posted within the geospatial boundaries of Baltimore City, MD from April 16, 2015 to May 4, 2015. These dates were

[1]While officers testified that they believed Gray's knife was illegal, the Maryland state attorney for Baltimore later clarified that Gray in fact was in possession of a spring-assisted knife that was legal under Maryland law [4].

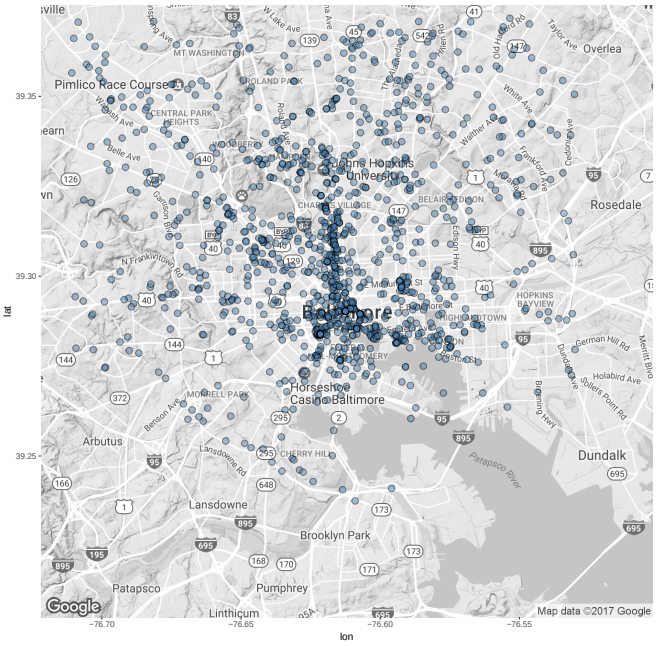

Figure 1: Geographical distribution of social activity hubs in Baltimore, MD estimated using the DPM-based local minima submodel. SAHs cluster in areas with high daytime populations like downtown Baltimore and the high population area around Johns Hopkins University. The process for estimating these locations is described in section 3.

selected because they were when the protests related to the death of Freddie Gray occurred in the city. Our sample included a total of 111,440 tweets made by 7,884 unique users. In order to restrict our analysis to people with a positive probability of protesting, we limited our sample to only include geotagged tweets. Present illustrations are based on smaller subsets of these users.

In order to estimate these users' SAHs, we collected up to 3,200 of each user's most recent tweets using Twitter's API, which we called through the TwitteR package in R [16].[2] Tweets were collected between July 19, 2016 and August 27, 2016. Each user's tweets are then narrowed to those that contain geotags. Each geotagged tweet in this final sample is treated as an observed incident of the user's movement patterns, and from the collection of all observed incidents, we estimate the user's SAHs.[3]

By default, a twitter user's location is not displayed when posting a 140 character message to twitter. However, users can identify their location when tweeting by enabling the location services that twitter provides. A user is able to selectively add location information, such as a geographic area (city or neighborhood), or a precise location in terms of latitude and longitude coordinates from the global positioning system that is available in most smart phones. Importantly, opting to share the location of a tweet is a social act. Because this is central to our research question, we do not include data from those individuals who chose not to disclose their geographic location.

3 ESTIMATING SOCIAL ACTIVITY HUBS

In this section, we present in detail the method we used to estimate Twitter users' SAHs. As the availability of information associated with each Twitter account differs, our SAH model, summarized in algorithm 1, is conditional on what this information affords, defaulting to more basic models where data availability is low. More specifically, we intend to define SAHs in two ways described in more detail below, based on posterior quantities of a Dirichlet process mixture (DPM) model for spatial data [35]. Estimation relies on an MCMC algorithm,[4] which is subject to convergence difficulties. In these rare cases, we document the specific user and return to diagnose potential issues. We discuss this phenomenon and our solutions more explicitly in section 3.1.1.

3.1 Dirichlet Process Mixture Model for Spatial Data

For users whose tweets contain sufficient information regarding their movement patterns, we use the Dirichlet process mixture (DPM) model of geographic profiling as the basis of our SAH model. DPM models for spatial data, based on prior geographic profiling models in criminology [27, 28], was first described in [35] where it was applied to spatial epidemiology. More recently, the model was used in an attempt to determine the identity of graffiti artist Banksy [20].

The intuition of the DPM model for spatial data is to sort a set of observed incidents in physical space into clusters originating from different source locations, without prior assumptions about the number of clusters that exist. For our present purposes, the DPM model is preferred over alternatives that require a fixed number of clusters (including those with a single cluster), because individuals are likely to vary in terms of their movement patterns (of which we have no prior data). Where there are multiple clusters, especially when they are highly dispersed, a misspecified number will result in inaccurate source location estimates that are skewed by "outliers," which are actually observations that originate from a different source.

The DPM model rectifies this by estimating the number of sources based on the observed data. The flexibility afforded by this feature is especially desirable, given the large number of Twitter accounts we are working with, as it is not feasible to adjust the SAH model for each Twitter account individually. A DPM model is not without assumptions, which are provided in the description below. In short, by employing the DPM model, we assume that individuals can have multiple SAHs from where their movement outward follows identical distributions, which in this implementation we specify as

[2] API Documentation available here: https://dev.twitter.com/rest/reference/get/statuses/user_timeline.

[3] As our research strategy enables us to estimate Twitter users' Social Activity Hub location, it is pertinent to address ethical concerns regarding the steps we take to maintain anonymity and protect users from potential harm. First, the sample is only taken from users who had opted into sharing their location with Twitter. By default, Twitter does not record the location where a tweet was posted. Instead, users must change their phone's settings to give Twitter permission to record their location via GPS. Second, we anonymize Twitter account names by applying a cryptographical hash. Third, the estimate standard deviation of the location means that we are only able to know the location of the Social Activity Hub within three miles.

[4] The algorithm is implemented in the Rgeoprofile package for R [34].

Algorithm 1: Social Activity Hub Estimation for Each User

Data: The set of n observed incidents $\mathbf{x}_i = (x_i^{(1)}, x_i^{(2)})$, $i = 1, ..., n$
if $n = 1$ **then**
  Assign $\mathbf{x}_1$ as $\text{SAH}^{(\text{sole observed})}$;
else
  MCMC algorithm implemented as the `RunMCMC()` function in Rgeoprofile 1.2 [34, summarized below in algorithm 2], based on the discussion in Verity et al. [35]);
  if *convergence fails;*
  then
    Document failure;
  else
    Take 3000 posterior draws; thin by keeping the first of every 30;
    begin local minima model:
      combine all 100 posterior draws;
      a) calculate hitscore surface;
      b) find local minima j on surface within σ degree decimal radius, $j = 1, ..., \infty$;
      c) **foreach** *local minimum j* **do**
        Assign $\mathbf{x}_i$ closest in Euclidean distance to local minimum as $\text{SAH}_j^{(\text{local minima})}$
      end
      foreach *posterior draw* **do**
        a to c;
      end
    end
    begin cluster mean model:
      combine all 100 posterior draws;
      d) **foreach** *cluster j of* $\mathbf{x}$ **do**
        Assign $\mathbf{x}_i$ closest in Euclidean distance to estimated source of cluster as $\text{SAH}_j^{(\text{cluster mean})}$
      end
      foreach *posterior draw* **do**
        d;
      end
    end
  end
end
Result: **SAH**=($\text{SAH}^{(\text{sole observed})}$, $\text{SAH}^{(\text{local minima})}$, $\text{SAH}^{(\text{cluster mean})}$)

a bivariate normal distribution, with standard deviation varying by specific application.

More specifically, the DPM model we use, adapted for spatial data by Verity et al. [35], is as follows. For each Twitter user, define a two-dimensional sample space with a finite grid of cells as Ω, in which each cell $\boldsymbol{\omega} = (\omega^{(1)}, \omega^{(2)})$ is a vector containing the latitude and longitude in decimal degrees of a geocoordinate. The set of n geocoordinates obtained from geotagged tweets $\mathbf{x} = \mathbf{x}_{1,...,n}$ is assumed to be the result of independent draws from a mixture of a countably infinite set of bivariate normal distributions centered on $\mathbf{z} = \mathbf{z}_{1,...,\infty}$, each with a variance of σ^2; (σ contains expectations about the movement patterns of individuals and must be specified by the user). Both $\mathbf{x}$ and $\mathbf{z}$ are defined on Ω. The prior distribution of the set of $\mathbf{z}$ is assumed to be a bivariate normal centered on the mean of $\mathbf{x}$, with a variance of τ^2 (τ is set to the largest distance in either longitude or latitude). c_i is a categorical variable that assigns $\mathbf{x}_i$ to source $\mathbf{z}_{c_i}$, and is drawn from a Dirichlet process, specifically the Chinese Restaurant Process which has a concentration parameter α drawn from a diffuse hyper-prior (specifically $h(\alpha) = ((1+\alpha)^2)^{-1}$) and a base distribution that is the bivariate normal (with mean $\mathbf{x}/n$) discussed above. This is formally represented as,

$$\begin{aligned} \mathbf{x}_i|\mathbf{z}_{c_i} &\sim \mathcal{N}(\mathbf{z}_{c_i}, \Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}) \\ \mathbf{z}_{1,...,\infty} &\sim \mathcal{N}(\mathbf{x}/n, \mathrm{T} = \begin{bmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{bmatrix}) \\ c_i &\sim CRP(\alpha) \\ \alpha &\sim \mathcal{H} \end{aligned} \tag{1}$$

Exact computation of posterior quantities are intractable when the number of observations is high ($n > 10$ being a useful rule of thumb; see [35] for analytical solutions to relevant posterior quantities), but can be estimated using MCMC methods [26, 35], which is implemented in the R package Rgeoprofile 1.2 [34]. The MCMC algorithm (`RunMCMC()` presented in algorithm 2) is based on a Gibbs sampler that alternates between draws of source locations $\mathbf{z}_{c_i}$ and cluster assignment c_i for all $i = 1, ..., n$ observations. The algorithm returns, for each $\mathbf{x}_i$, its cluster c_i; and for each unique cluster c_j, its spatial mean $\mathbf{z}_j$.

Algorithm 2: `RunMCMC` from Rgeoprofile 1.2 [34]

Data: The set of n observed incidents $\mathbf{x}_i = (x_i^{(1)}, x_i^{(2)})$, $i = 1, ..., n$
Initialize by setting initial values and computing relevant priors;
Define sampling steps:
  a) draw and update $\mathbf{z}_{c_i}$ based on most updated c_i;
  b) draw and update c_i based on most updated $\mathbf{z}$;
begin Burn-in
  repeat
    for *i in* 1 *to n* **do** a-b;
  until *convergence*;
end
begin Posterior draws
  foreach *posterior draw* **do**
    for *i in* 1 *to n* **do** a-b;
  end
end
Result:
  1. For each $\mathbf{x}_{1,...,n}$, its corresponding cluster c_i
  2. For each unique cluster c_j, its source location $\mathbf{z}_j$

3.1.1 Diagnosing Convergence Difficulties. MCMC convergence is assessed by the potential scale reduction factor (psrf) evaluated on the log-likelihood of the model [15]. This assessment is implemented as `gelman.diag()` the in the coda package. MCMC chains are taken to have converged when the upper bound of the psrf falls below 1.1 following a burn-in period, which we specified as 300 draws. Generally, models successfully converge within the burn-in period or shortly after, but one of two issues may arise.

First, a small number of models take a disproportionally long time to reach convergence. This usually occurs for users with a large set of observed incidents. Because we have a need to estimate a large number of SAHs, we specify a maximum burn-in of 3,000 iterations, at which point if the psrf is not below 1.1, the specific user is documented for manual diagnosis. For reference, consider that in the two illustrations presented below, failure to converge after 3,000 burn-in draws occurred 14 times out of 200 users estimated, and one time out of 126 users estimated. In fact, for most data sets, convergence was achieved within the minimum burn-in period of 300 or shortly after. For the present illustrations, we drop users who do not converge after 3,000 burn-in draws from our examination as representative sampling is not a requirement.

Second, extreme sparsity in data in the form of singular observations or spatially dispersed observations without overlap (which are singular observations within certain clusters) can result in an error when computing the psrf. To see why this occurs, first note that the log-likelihood of the model is calculated based on the fit of the distribution of observed incidents into clusters, including the number of clusters present and how the observed incidents are sorted among them. This occurs after sampling step b in algorithm 2. In instances of high data sparsity and dispersion, cluster assignments c_i are never updated through sampling step b because the probability of assigning a different cluster, which is updated in step a, while always nonzero, is extremely low. The result is that for all MCMC iterations across all chains, the same log-likelihood is computed based on the unchanging distribution of observed incidents into clusters. In short, the Gibbs sampler immediately moves to a very small area and any movement within this area does not yield probabilities for different clustering combinations that is meaningfully above zero. The log-likelihood which is computed based on this clustering therefore remains constant, leading to errors when attempting to compute the psrf as it is based on variation within and across MCMC chains. We document these errors, but take the modeled results as the best estimate of SAHs given the available information, and as such, take these models as having converged. The most extreme case is where there is only a singular observation, which we immediately take to be the SAH as outlined in algorithm 1.

3.2 Local Minima and Cluster Mean Submodels

As introduced earlier, we use the posterior quantities obtained from the DPM model in our SAH model in two ways. For the *local minima* submodel, begin by defining $\mathbf{S} \subseteq \Omega$ as the grid bound by the minimum and maximum values of the set of observed $\mathbf{x}$. Next, for every cell $\mathbf{s} \in \mathbf{S}$, rank $\mathbf{s}$ according to the sum of its distances to each source location $\mathbf{z}_j$ over all posterior draws, where distance is not linear but weighed by the inverse of the bivariate normal density around $\mathbf{z}_j$. Consistent with existing geoprofiling approaches [e.g. 28], ranks are transformed to hitscores on $[0, 1)$, but remain functionally equivalent in that lower is better and all values are distinct.[5] This type of hitscore surface is traditionally used as a surface for search priority of source locations [28, 35]. On this surface, we find all m local minima (i.e. locations with higher priority) within an approximately two mile radius (0.05 decimal degrees) and define a user's SAHs as the set of m observed $\mathbf{x}_i$ closest to these local minima. For the *cluster mean* submodel, we define a user's SAHs as the set of observed $\mathbf{x}_i$ closest to the set of estimated source locations $\mathbf{z}$ averaged across posterior draws.

Figure 2 illustrates the SAHs estimated under both submodels in relation to the hitscore surface produced by the DPM model. For this Twitter account, the DPM model aggregated over all posterior draws estimated the set of observed incidents to have originated from two sources (i.e. $\mathbf{z}_j, j = 1, 2$). As evident from Figure 2, the two submodels agreed on a potential source $\mathbf{z}_1$ in the upper left of the physical space (directly north of Baltimore City) as an SAH. In the bottom left (directly west of Baltimore City), observations are not dispersed enough to consistently yield a third cluster, but are relatively sparse, such that the estimates for the second source varied greatly. In fact, between different posterior draws, the DPM model assigned the set of $\mathbf{x}$ not associated with $\mathbf{z}_1$ to either $\mathbf{z}_2$ or $\mathbf{z}_3$. Because of this, the difference between the local minima and cluster mean submodels (based on how posterior draws are aggregated, i.e., sum of computed probabilities versus means), leads to disagreement between the submodels on the second SAH. This example illustrates the importance of understanding uncertainty in the DPM model, which we discuss next.

3.3 Uncertainty in the DPM Model

In earlier applications of the DPM model to spatial data, there is justifiably less of a concern over the uncertainty of estimates. However uncertain, the expected values of $\mathbf{z}$ are what informs a search that must take place. Existing implementations of the model [34, e.g.,] therefore do not readily yield uncertainty measures. However for inferential modeling, measures of uncertainty feature much more prominently. In order to account for uncertainty in our SAH model, we take 3,000 draws from the posterior distribution of the DPM model, thinned to 100 samples, and use this information to determine a set of corresponding SAHs following both the local minima and cluster mean submodels. Specifically, for the local minima submodel, instead of computing a hitscore surface based on all posterior draws, we do so for each draw independently; and for the cluster mean submodel, $\mathbf{z}$ is not averaged across posterior draws. SAH estimates are stored for each posterior draw, forming a posterior distribution of SAHs. This distribution can be used in subsequent statistical modeling to account for uncertainty associated with the SAH model. In the remainder of this section, we use the same Twitter account as above to illustrate uncertainty within the two DPM-based submodels. This particular account was chosen because it is illustrative both in terms of its estimates and the uncertainty associated with them. The level of uncertainty associated

[5] The two computational steps above are implemented in the `ThinandAnalyse()` function in Rgeoprofile 1.2 [34].

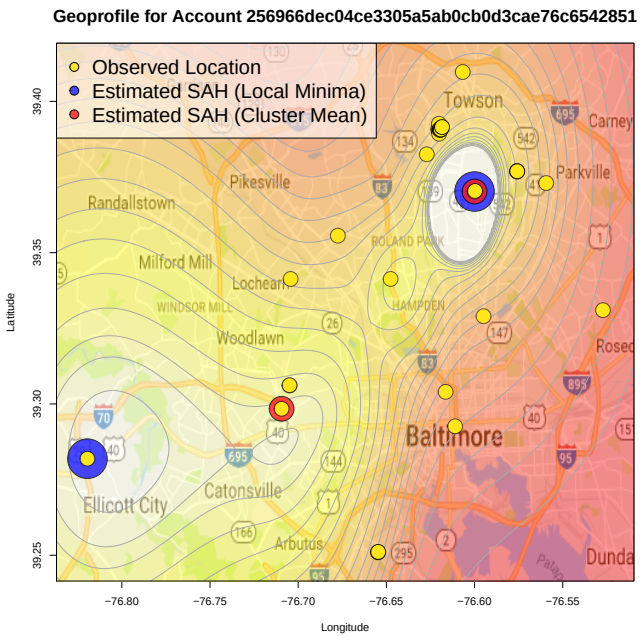

Figure 2: Example of a single user's estimated social activity hubs in relation to the hitscore surface produced by the DPM model. Yellow points are observed incidents. Points enclosed in blue indicate SAHs determined by the local minima submodel. Points in enclosed in red indicate those determined by the cluster means submodel.

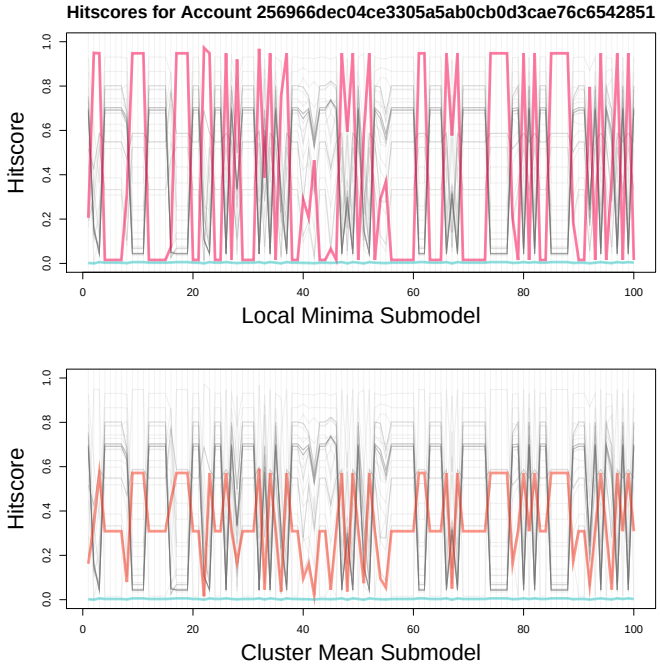

Figure 3: Visualization of uncertainty about SAH estimates. Each line indicates a source location and its variation in hitscore across different draws. Highlighted lines indicate source locations selected by the algorithm as an SAH.

with this account, based on visual inspection, is neither particularly high nor low.

In order to visualize the level of uncertainty about our SAH estimates, we plot the variation in hitscore associated with each potential source across all posterior draws. Specifically, in Figure 3 (which corresponds to the hitscore surface in Figure 2), each horizontal line documents the hitscore of a potential source location as it varies across posterior draws. The highlighted lines are the sources chosen as the SAHs from the combined posterior draws, which may differ depending on the submodel used. The highlighting color serves to tie corresponding SAHs across the two subfigures. Variation in the hitscore indicates changes in the topography of the hitscore map across posterior draws. This does not, however, necessarily mean that there is uncertainty about the SAH estimates, which arise when changes in the topography are large enough to induce changes in the hitscore rank of potential sources relative to one other. This uncertainty is indicated by lines that cross. Future work may benefit from a formal quantification of this type of uncertainty.

As discussed earlier, the DPM model estimated the set of observed incidents associated with this account to have originated from two sources. As is shown in Figure 3, corroborating Figure 2, both DPM-based submodels yielded the potential source location with the lowest hitscore as one of the two SAHs. This SAH_1, highlighted in turquoise, is associated with very little uncertainty. The variation in its hitscore across all posterior draws is minor, and it maintains a stable hitscore rank in all but one draw (42). On the other hand, the second SAH_2, highlighted in two shades of coral, differs between the two submodels and exhibits high levels of uncertainty within each submodel. As is evident from the figure, the local minima $SAH^{(localminima)}$ varies widely between posterior draws, but for the majority of these draws, it is the best performing SAH (in terms of hitscore), just after the SAH_1. The cluster mean $SAH^{(clustermean)}$ is not selected based on the hitscore of potential source locations, but still yields a relatively desirable result according to this criterion; while its hitscore based on the aggregated posterior draws is not as low as that of the $SAH_2^{(localminima)}$, it is subject to less of the extreme fluctuation across draws exhibited by the local minima submodel.

At the same time, the present discussion serves to illustrate the potential for high levels of uncertainty to be associated with DPM-based SAHs. Researchers intending to utilize these estimated quantities should be mindful of this when making statistical inferences. As we proposed, the entire posterior draw can be used in statistical modeling, which allows for the construction of confidence bounds. Our novel method paves the way future work in formal quantification of the uncertainty discussed here and in better understanding the inferential shortcomings associated with not properly accounting for this source of uncertainty.

4 ILLUSTRATIONS

In this section, we illustrate the validity of our measurement algorithm in two ways, as comparison to traditional spatial estimation models, and based on its predictive validity.

4.1 Comparison to Spatial Means

First, we compare for each user the performance of their SAHs obtained from the local minima and cluster mean submodels relative to the source closest to the spatial mean of their observed incidents. To do so, we randomly selected 200 users from the sample described in section 2.2. Users whose DPM model did not converge were dropped as discussed in section 3.1.1, yielding a final sample of 186. For each user, we estimated their $\text{SAH}^{(localminima)}$ and $\text{SAH}^{(clustermean)}$.[6] We also calculated the spatial mean of their observed incidents, and similar to how we estimate SAHs, specify the potential source closest to the spatial mean as the estimated source. Then, we calculate the Euclidean distance between each of the user's observed incidents to its closest $\text{SAH}^{(localminima)}$, $\text{SAH}^{(clustermean)}$, and the spatial mean-estimated source. Finally, we record the median distance for each estimation method. We repeated these steps for each of the 186 users. The results are summarized in Table 1 as percentiles of the median distances for each estimation method.

Distance between Estimated Sources and Incidents

	Percentiles				
Method	5	25	50	75	95
Local Minima	0.00	0.0005	0.02	0.07	0.40
Cluster Mean	0.00	0.0004	0.01	0.02	0.05
Spatial Mean	0.00	0.0651	0.33	1.67	20.22

$n = 186$

Table 1: Comparison of different estimation methods. Cells are the percentile values of the median distance in decimal degrees from each user's observed incidents to their closest SAH/spatial mean.

It is evident from the results that the source estimated based on the spatial mean tends to be considerably more distant to the user's observed incidents. Based on this, we believe that the face validity of the spatial mean is generally low. These results highlight the importance of a data-driven approach to estimating the optimal number of clusters in social media data. Forcing data into a user-determined number of clusters (one in this example) biases estimated movement patterns, which may critically influence subsequent inferential steps.

4.2 Predictive Validity

Next, we assess the predictive validity of our SAH measures. To do so, we compare whether the SAHs from known protesters in our data are more likely to fall within known protest locations compared to a sample of users whose protest behavior is unknown.

After drawing a random subset of 1,000 tweets from the sample described in section 2.2, we hand-code whether each tweet indicated attendance in the Freddie Gray protests. We identify 64 unique protesters within this subset. From the same subset, we randomly select 62 users who we did not identify as protesters. Then, we estimate our two SAH measures for all 126 users.[7]

Using the resulting SAHs, we next consider whether there are different geospatial patterns in the SAHs from protesters and non-protesters. Specifically, we assess the proportion of users from both samples with at least one SAH within a known Freddie Gray protest location. We identify these locations based on data from Baltimore news sources published in April and May, 2015. These locations are summarized in Table 2.

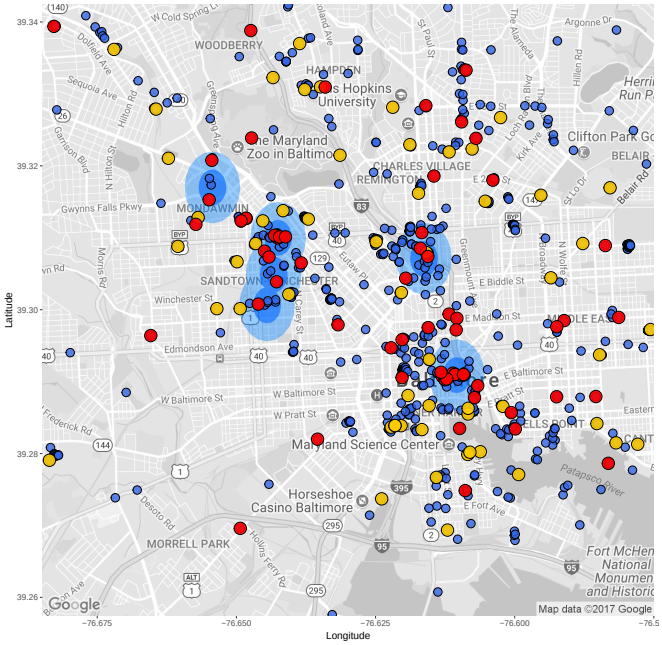

Figure 4: Visualization of results from the validation exercise to assess whether SAHs recover different patterns among protesters and non-protesters. The six light blue circles indicate known protest locations. Dark blue dots are all observed tweets. Red dots are SAHs that belong to protesters, while yellow ones belong to non-protesters. As the figure show, protesters' SAHs tend to be closer to known protest locations. These results are presented in Table 3.

Results from our examination are presented in Table 3. As these results indicate, there is a stark contrast between the SAH of users from whom we have observed protest behavior and those for whom we have no information. These results demonstrate the ability of our measure to capture meaningful behavior patterns based on observed movements, further supporting the validity of our method.

[6] Model parameters are specified to be the ones presented: $\sigma = 0.05$ decimal degree; minimum burn-in of 300; maximum burn-in of 3,000; 3,000 posterior draws thinned to 100. Convergence results are: 14 did not converge after 3,000 burn-in draws. Data for 42 users were at least conditionally sparse in observations, with 9 of these being single observations.

[7] Because the aim is to identify localized movement patterns as opposed to more general hubs, model parameters are specified slightly differently: $\sigma = 0.01$ decimal degree; minimum burn-in of 300; maximum burn-in of 3,000; 3,000 posterior draws thinned to 100. Convergence results are: 1 did not converge after 3,000 burn-in draws. Data for 24 users were at least conditionally sparse in observations, with 10 of these being single observations.

Locations of Freddie Gray Protests

Location	Longitude	Latitude
North and Penn	-76.6425	39.3100
Baltimore Police Dept. West	-76.6445	39.3006
Baltimore City Hall	-76.6104	39.2909
Gilmor House	-76.6433	39.3049
Mondawmin Mall	-76.6543	39.3170
Penn Station	-76.6163	39.3071

Table 2: Summary of locations in or near Baltimore City bounds specified as a protest location during the Freddie Gray protests, April-May, 2015.

Proportion of Users with SAHs in Protest Locations

	Local Minima	Cluster Mean
Protesters	0.37	0.27
Random Sample	0.02	0.02
Difference	0.35	0.25

$n_{protest} = 63, n_{random} = 62$

Table 3: Summary of SAH predictive validity. Cells are the proportion of users from either sample with at least one estimated SAH falling within a known protest location. Protest location is defined by a 0.0025 decimal degree radius around each of the six coordinates specified in Table 2. Using a 0.005 decimal degree radius yields similar results.

5 DISCUSSION

In this paper, we contribute to the developing field of research on inferring geospatial context from volunteered geographic information, and presented a method to estimate Twitter users' "social activity hubs" (SAHs), or the geospatial areas where users spend time throughout the day. As a validation exercise, we linked these locations to incidences of political participation, in particular the protests that transpired over the death of Freddie Gray, in Baltimore during April and May of 2015. The patterns we discovered suggest the methods proposed here for estimating SAHs are able capture meaningful measures of sociopolitical context.

ACKNOWLEDGMENTS

The project has been reviewed and approved by the University of Michigan Institutional Review Board (HUM00126649). It was determined to not be human research by the Pennsylvania State University Institutional Review Board (STUDY00007059). We acknowledge research support from the Charles Koch Foundation, Emory's Institute for Quantitative Theory and Methods, the National Science Foundation's XSEDE program (SES170013), and the San Diego Supercomputer Center. We thank Kevin Reuning, four anonymous reviewers, and conference attendees at MPSA, PolNet, and PolMeth for helpful feedback and suggestions.

REFERENCES

[1] Maria Abascal. 2015. Us and them: Black-White relations in the wake of Hispanic population growth. *American Sociological Review* 80, 4 (2015), 789–813.
[2] Pablo Barberá. 2014. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis* 23, 1 (2014), 76–91.
[3] Matt A Barreto, Gary M Segura, and Nathan D Woods. 2004. The mobilizing effect of majority–minority districts on Latino turnout. *American Political Science Review* 98, 1 (2004), 65–75.
[4] Alan Blinder and Richard Pérez-Pe na. 2015. 6 Baltimore Police Officers Charged in Freddie Gray Death. *New York Times* (May 2015).
[5] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D.I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. A 61-Million-Person Experiment in Social Influence and Political Mobilization. *Nature* 489, 7415 (2012), 295–298.
[6] David Broockman and Joshua Kalla. 2016. Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science* 352, 6282 (2016), 220–224.
[7] Joseph Brown, Gordon McCord, and Paul Zachary. 2017. Sunday, Bloody Sunday: Evidence from Northern Ireland for the Effect of Ethnic Diversity on Violence. (June 2017). Working paper.
[8] Ted Hsuan Yun Chen, Paul Zachary, and Christopher J. Fariss. 2017. Who Protests? Using Social Media Data to Estimate How Social Context Affects Political Behavior. (October 2017). Working Paper.
[9] Kyle Crowder. 2000. The racial context of white mobility: An individual-level assessment of the white flight hypothesis. *Social Science Research* 29, 2 (2000), 223–257.
[10] Ryan D Enos. 2014. Causal effect of intergroup contact on exclusionary attitudes. *Proceedings of the National Academy of Sciences* 111, 10 (2014), 3699–3704.
[11] Ryan D Enos. 2016. What the demolition of public housing teaches us about the impact of racial threat on political behavior. *American Journal of Political Science* 60, 1 (2016), 123–142.
[12] Ryan D Enos and Noam Gidron. 2016. Intergroup behavioral strategies as contextually determined: Experimental evidence from Israel. *The Journal of Politics* 78, 3 (2016), 851–867.
[13] Christopher J. Fariss and Zachary M. Jones. 2017. Enhancing Validity in Observational Settings When Replication is Not Possible. *Political Science Research and Methods* https://doi.org/10.1017/psrm.2017.5 (2017).
[14] Laura K. Gee, Jason J. Jones, Christopher J. Fariss, Moira Burke, and James H. Fowler. 2017. The Paradox of Weak Ties in 55 Countries. *Journal of Economic Behavior and Organization* 133, Janurary (2017), 362–372.
[15] Andrew Gelman and Donald B Rubin. 1992. Inference from iterative simulation using multiple sequences. *Statistical science* (1992), 457–472.
[16] Jeff Gentry. 2015. *twitteR: R Based Twitter Client.* http://CRAN.R-project.org/package=twitteR R package version 1.1.9.
[17] Anna Getmansky and Thomas Zeitzoff. 2014. Terrorism and voting: The effect of rocket threat on voting in Israeli elections. *American Political Science Review* 108, 3 (2014), 588–604.
[18] Rob Grace, Jess Kropczynski, Scott Pezanowski, Shane Halse, Prasanna Umar, and Andrea Tapia. 2017. Social Triangulation: A new method to identify local citizens using social media and their local information curation behaviors. In *Proceedings of the 14th ISCRAM Conference.* 902–915.
[19] Samiul Hasan, Xianyuan Zhan, and Satish V Ukkusuri. 2013. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing.* ACM, 6.
[20] Michelle V Hauge, Mark D Stevenson, D Kim Rossmo, and Steven C Le Comber. 2016. Tagging Banksy: Using geographic profiling to investigate a modern art mystery. *Journal of Spatial Science* 61, 1 (2016), 185–190.
[21] Bin Jiang and Jean-Claude Thill. 2015. Volunteered Geographic Information: Towards the establishment of a new paradigm. *Computers, Environments and Urban Systems* 53 (2015), 1–3.
[22] Gary King. 2013. *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data.* Princeton University Press, Princeton, NJ.
[23] Kevin Kruse. 2005. *White Flight: Atlanta and the Making of Modern Conservatism.* Princeton University Press, Princeton, New Jersey.
[24] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas A. Christakis, Noshir Contractor, James H. Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational Social Science. *Science* 323 (2009), 721–723.
[25] David Lazer and Jason Radford. 2017. Data ex Machina: Introduction to Big Data. *Annual Review of Sociology* (2017).

[26] Radford M Neal. 2000. Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics* 9, 2 (2000), 249–265.
[27] Mike O'Leary. 2010. Implementing a Bayesian approach to criminal geographic profiling.. In *COM. Geo*.
[28] D Kim Rossmo. 1999. *Geographic profiling*. CRC press.
[29] Melissa L Sands. 2017. Exposure to inequality affects support for redistribution. *Proceedings of the National Academy of Sciences* (2017), 201615010.
[30] William R. Shadish. 2010. Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings. *Psychological Methods* 12, 1 (2010), 3–17.
[31] William R. Shadish, Thomas D. Cook, and Donald T. Campbell. 2001. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth Publishing.
[32] Zachary C. Steinert-Threlkeld. 2017. Spontaneous Collective Action: Peripheral Mobilization During the Arab Spring. *American Political Science Review* (2017).
[33] Zachary C. Steinert-Threlkeld, Delia Mocanu, Alessandro Vespignani, and James H. Fowler. 2015. Online Social Networks and Offline Protest. *EPJ Data Science* 4, 19 (2015), 1–9.
[34] Stevenson, M.D, Verity, and R. 2014. *Rgeoprofile : Geographic Profiling in R.* Queen Mary University of London, London, England. http://evolve.sbcs.qmul.ac.uk/lecomber/sample-page/geographic-profiling/ Version 1.2.
[35] Robert Verity, Mark D Stevenson, D Kim Rossmo, Richard A Nichols, and Steven C Le Comber. 2014. Spatial targeting of infectious disease control: identifying multiple, unknown sources. *Methods in Ecology and Evolution* 5, 7 (2014), 647–655.