Gang Networks, Neighborhoods and Holidays: Spatiotemporal Patterns in Social Media

Nibir Bora*, Vladimir Zaytsev*, P. Jeffrey Brantingham†, Yu-Han Chang* and Rajiv Maheswaran*
*Information Sciences Institute, University of Southern California, Marina del Rey, CA 90292
†Department of Anthropology, University of California Los Angeles, Los Angeles, CA 90095
Email: {nbora,zaytsev}@usc.edu, branting@ucla.edu, {ychang,maheswar}@isi.edu

***Abstract*—Social media generated by location-services-enabled cellular devices produce enormous amounts of location-based content. Spatiotemporal analysis of such data facilitate new ways of modeling human behavior and mobility patterns. In this paper, we use over 10 millions geo-tagged tweets from the city of Los Angeles as observations of human movement and apply them to understand the relationships of geographical regions, neighborhoods and gang territories. Using a graph based-representation of street gang territories as vertices and interactions between them as edges, we train a machine learning classifier to tell apart rival and non-rival links. We correctly identify 89% of the true rivalry network, which beats a standard baseline by about 30%. Looking at larger neighborhoods, we were able to show that distance traveled from home follows a power-law distribution, and the direction of displacement, i.e, the distribution of movement direction, can be used as a profile to identify physical (or geographic) barriers when it is not uniform. Finally, considering the temporal dimension of tweets, we detect events taking place around the city by identifying irregularities in tweeting patterns.**

I. Introduction

With the proliferation of social media and Global Positioning System (GPS) enabled cellular devices, and the subsequent habituated sharing of geo-location information in various kinds of Internet footprints (e.g., tweets[1], check-ins[2]), creates a valuable sources of real-time data to monitor and model human behavior and mobility patterns. Human mobility is characterized by a number of subtle interesting phenomena, for instance the process of traveling to work, visiting friends, buying groceries, recreation, etc. These phenomena contain patterns. Social and geographic landscapes of districts and neighborhoods, and political and physical barriers induce interactions between humans leading to the formation of groups, competition between groups, power struggles, conflict, alienation that lead to mobility patterns. Discovering these patterns and their underlying explanations attracts researchers to study and model these behaviors.

Recently, statistical models have been used to understand spatiotemporal elements of human behavior. Daily travel patterns tend to show bursty and heavy tailed behavior [1]. Neighborhoods and street gang territories and known to evolve over time and need not be as strict as initially drawn by city planners [2]. Burglaries and similar criminal activities tend to be repeated in close vicinities [3]. Turfs between street gangs and retaliatory behavior show much resemblance to earthquakes and aftermath [4]. These characteristics have been observed and validated using archival data collected by scientists, reporters and law enforcement. Another interesting phenomena in social media is the breakout of news and updates regarding important events. The modern paradigm of social media sharing and re-sharing allows such news to spread rapidly, allowing tracking on a geographic space when the media is geo-tagged.

In this paper, we attempt to model human behavior patterns and social interactions using strictly geo-tagged data from the city of Los Angeles, California, collected using Twitter's streaming service. Following a user-centered model, we identify unique users and their home locations using their geo-location data shared by them while tweeting. Coupled with social boundaries like regions and neighborhood, we analyze their activity on a geographic canvas. Our first experiment pertains to predicting rivalries between street gangs of the Hollenbeck policing district of eastern Los Angeles. Modeling the gangs territories as a graph with edges showing visits by people from one territory to another, we use a machine learning classifier to predict if these links are of rival or non-rival nature, thus recovering a rivalry network between the gangs which can be validated against the actual rivalry network know to experts. On the geographic landscape of the city, we show that the spatial distribution of direction traveled by users in a region, as estimated by social media can be used to discover obstructions such physical barriers like freeways, coastlines and mountains. Looking at neighborhoods in any region in the city, we analyze the entropy of incoming and outgoing flows to see if we can characterize important hubs. Finally, we analyze the temporal dimension to the geo-location data and see if an aberration from a baseline activity pattern can be used to identify events occurring in different regions of the city.

The next section introduces related work about spatiotemporal analysis of mobility and human activity. Section III describes the data that was collected and how it was prepared to be used in the experiments. Section IV details the process of modeling interactions among street gangs and predicting a rivalry network between them. Section V is a series of qualitative and quantitative observations in the data which can be used to model travel patterns of people. Section VI talks about identifying events in different regions of Los Angeles by identifying skews in activity pattern of people. We summarize our conclusions in Section VII and talk about future directions.

[1] http://twitter.com
[2] http://foursquare.com

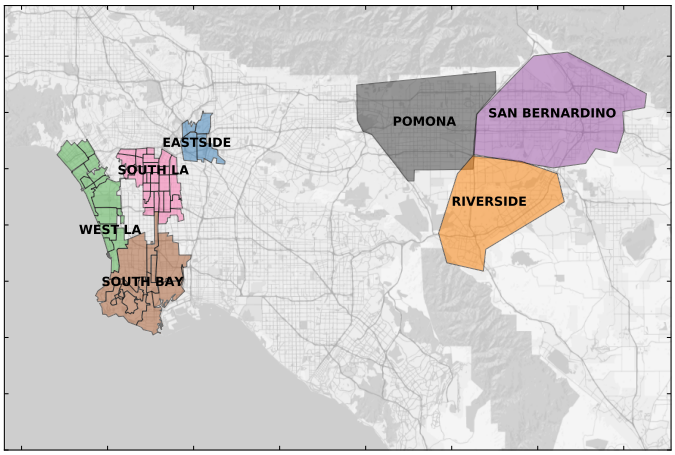


Fig. 1: Figure shows map of the city of Los Angeles with region boundaries marked for few regions.

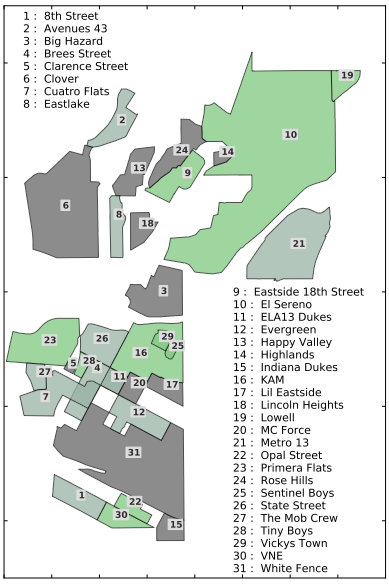


Fig. 2: Figure shows map of Hollenbeck policing district with territory boundaries marked for 31 street gangs.

II. Related Work

Temporal analysis of human activities like replying to emails, placing phone calls, etc. have shown that they occur in rapid successions of short duration followed by long inactive separations. Using a dataset of thousands of emails between users, the time interval between subsequent emails has been show to be of heavy-tailed nature, resembling a Pareto distribution [5]. Spatiotemporal models of human mobility have been studied on different datasets, such as circulation of US bank notes [6] and cell phone logs [1]. The truncated power law distribution characterizing heavy-tailed behavior for both the distance and time duration of hops between subsequent events in a trajectory of regular travel patterns has been established by many studies [6], [7], [1], [8], [9].

For geographic studies, user-generated geo-location-based data exposes a variety of avenues for spatiotemporal analysis. For example, researchers used 18 million geo-location-based check-in instances to shows the formation of 'live-hoods' from the social dynamics of people living in a city [2]. Live-hoods are pseudo-neighborhoods which do not necessarily correspond to actual neighborhood boundaries, but originate from them and in the course of time undergo transformation due to various demographic factors. Using a similar technique, researchers tried to perform cluster analysis to identify street gang territory structures by combining both social and geographical information gathered from Los Angeles Police Department (LAPD) Field Interview (FI) cards, corresponding to stopping of suspected gang members by patrolling officers [10], [11]. Even in the case of street gangs, the ecology of territories are dynamic and keep evolving over time due to constant clashes while competing for resources [12]. Researchers also validates ecological models of gang territorial patterns and territorial boundary patterns [13], [12]. From a social science perspective, studies like [14] try to understand the effect of geographical and environmental characteristics on territory boundaries, intervention experiments to reduce criminal violence, etc.

Several studies investigated the street gangs of Hollenbeck district in eastern Los Angeles. Most of these gangs engage in acts of violence over turf and respect, usually in the form of attacks against members of other gangs or infiltrating rival territory [15]. Researchers modeled the temporal dynamics of criminal activities between rival gangs and corresponding retaliatory behavior using self-exciting point process models (the Hawkes process), typically used to model earthquakes and aftershock events [4], [16] . While point process models adequately illustrates the temporal dynamics of such criminal patterns, [17] investigated an agent-based model (also known as the random walker program) which covers both the spatial and temporal aspect of rival crime patterns. The agent-based approach, which has been used to model human and animal behavior, simulates agents (criminals) as members of groups (gangs) by probabilistically modeling four processes: the movement of gang members in space, the addition of new gang members, fights between gang members, and updating of the rivalry strengths. Researchers validated both models using a dataset of crime events in Hollenbeck between 1999 and 2002 [17].

Modeling of criminal behavior has until now been performed using archival data, which sometimes might be inaccurate. Using real-time ubiquitous data collected from geo-location based sharing services like Twitter, such models of criminal behavior and gang activity can overcome the problems of using a static model in a largely dynamic social situation. Although geo-location data from Twitter has been used in various studies, for example spotting and tracking major events such as earthquake shakes [18], [19] or car accidents [20], so far it hasn't been used to model rivalry activity in street gangs. In [21], researchers used temporal and text features from tweets collected during three soccer games to identify game events in them. The key distinction in our study is that we discard the textual dimension of tweets and relies purely spatiotemporal information.

III. Data description

The data used in our experiments originates from a corpus of tweets collected from October 2012 to April 2013. We collected strictly geo-tagged tweets using Twitter's Streaming API[3] and limited them to the polygon bounding the entire city of Los Angeles. This way, we received all tweets with location information and not just a subset, without running into any rate limits. Each tweet is up to 140 characters of text and is associated with a user id, timestamp, latitude and longitude. We disregard all other fields obtained from Twitter, including the user's twitter handle, and in no way use the information we retain to identify personal information about the user. During the collection period, we accumulated approximately 10 million geo-tagged tweets.

To model social interactions and mobility, we begin by defining social boundaries on the geographic space. The city of Los Angeles is divided into several regions, each of which consist of a number of smaller neighborhoods. We adapt these region boundaries and neighborhoods from The Los Angeles Times - LA Mapping project[4] which provides an open API[5] to access all of their geographic data (16 regions and 114 neighborhoods). Figure 1 shows some of the region boundaries interpolated on a map of Los Angeles. Later, we will analyze patterns of mobility within these regions, and for that purpose pick a handful of interesting regions, namely, Eastside, Pomona, Riverside, San Bernardino, South Bay, South LA and West LA. The region Eastside spans a 15 square mile policing district called Hollenbeck which is home to 31 street gangs, making it of particular interest. Figure 2 shows each of the gang territories in the Hollenbeck district. The Mapping LA project does not account for these street gang territories, and hence we adapt them from [13] and [14].

Having defined the geographic scaffolding for our experiments, in order to model mobility, we now require user agents and their home locations. Although Twitter identifies its users by a unique identification, it does not account for the user's home location. Identifying home locations for users can be done using a very straight forward method, based on the assumptions that users generally tweet from home during the night. For each unique user tweeting in Los Angeles, we start by collecting all tweets between 7:00pm and 4:00am, and apply a single pass of DBSCAN clustering algorithm [22]. The largest cluster produced by the cluster analysis is then chosen as the one corresponding to the user's home location. The centroid of this cluster is used as the exact coordinates for the user's home. Skipping users with very few tweets, and ones for whom a cluster could not be formed, we we able to identify home locations for over 10,000 unique users. A user is assigned to a particular region or a neighborhood if his/her home coordinate pair lies within its geographic bounds.

After performing this preprocessing, we now have a rich set of data comprising of users, their home area, and their tweeting activity on a geographic space, which will allow us to intricately look into user mobility and interaction patterns. At this point, we discard the text element of each tweet since we do not perform any natural language processing on them. Retaining only the user identification, timestamp, latitude and longitude, we henceforth refer to these truncated tweets as *activity points*. This data will act as the seed for all our following experiments. We start off with an experiment on the interactions of users in Hollenbeck's street gang territories, and will then move on to understand mobility patterns in other regions and on a wider geographic canvas.

[3]https://dev.twitter.com/

[4]http://projects.latimes.com/mapping-la/

[5]http://projects.latimes.com/mapping-la/api/

IV. Predicting Gang Rivalry Network

Although studies on criminal activity patterns and street gang behavior are well established, they often rely on archival data from police and law keepers, which is often static and sometimes archaic. Street gangs have turned to Twitter to demean rivals and plan fights, making Twitter a promising real-time source of data for modeling gang behavior[6]. We will now attempt to predict the rivalry network between the street gangs of Hollenbeck (as in Figure 2) using the data we collected from twitter.

A. The Task

We define a complete graph G representing each street gang as a node (vertex) and links (edges) between each pair of gang territories. A link can either be a rival link or a non-rival link. The problem is then modeled as a classification task where the instances for classification are the links between street gangs. Attributing each of these links with features characterizing user activity patterns, we can train a machine learning classifier on it.

We start by assigning users to each of the street gang's territories. A user belongs to a particular street gang's territory if his/her home location lies within its bounds. It is important to point out that a user need not necessarily be an active member of the street gang just because his/her home location is in its territory. However our objective is to study the movement of people (gang member or not) in the presence of street gang territories. For n gangs $g_1, g_2, .., g_n$ in consideration, we define sets $\tau_{i,j}$, $i, j = 1, 2, .., n$, which consists of all activity points geographically lying within the territory of gang g_j, and the corresponding user belongs to gang g_i.

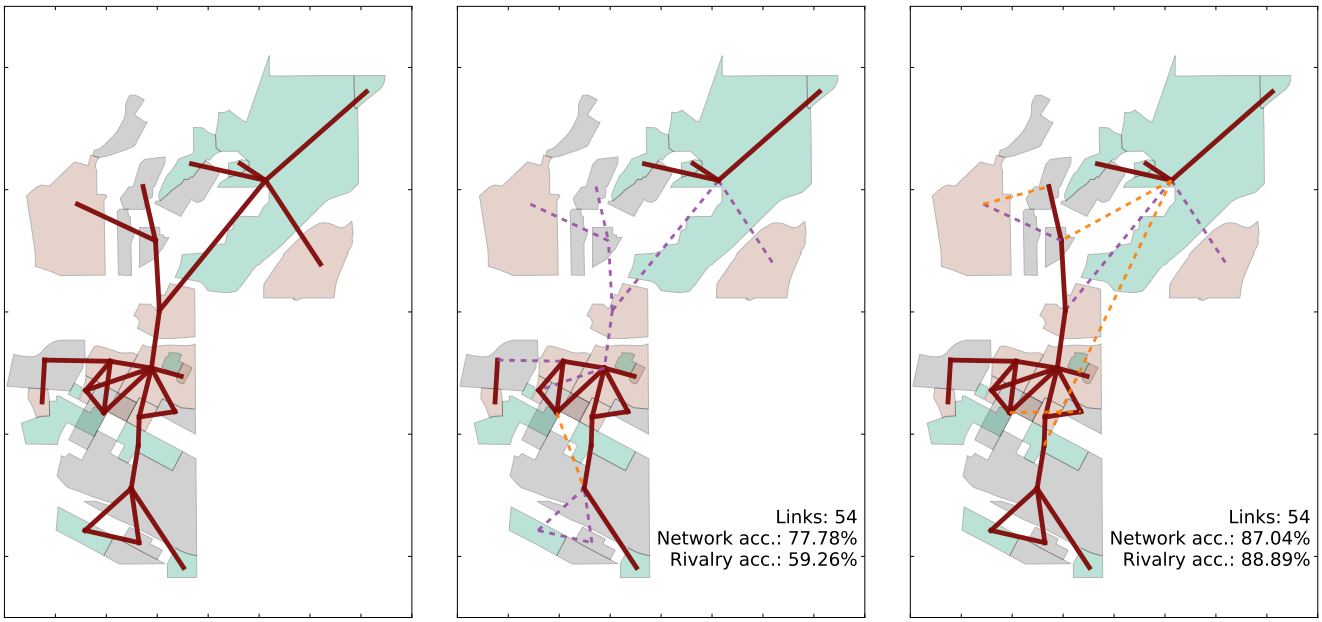
$$\tau_{i,j} = \{\text{tweets by users from } g_i \text{ in } g_j\text{'s territory}\} \quad (1)$$

The *activity matrix* $\mathbf{A_{i,j}}$ is a $n \times n$ matrix for gangs $g_1, g_2, .., g_n$ where each element is the number of activity points geographically lying within the territory of gang g_j, for which the corresponding user belongs to gang g_i, i.e., the number of elements in the set $\tau_{i,j}$.

$$\mathbf{A_{i,j}} = |\tau_{i,j}| \quad (2)$$

The activity matrix is a simple metric which designates interactions and mobility behavior of users belonging to the gang territories. A basic intuition would be that users from a particular gang territory generally stays away from their

[6]http://www.nydailynews.com/news/crime/gangs-new-york-talk-twitter-tweets-trash-talk-rivals-plan-fights-article-1.414083



(a) Actual rivalry network. (b) Baseline prediction. (c) Classifier prediction.

Fig. 3: Figure shows (a) Actual rivalry network (trimmed), (b) Rebuilt rivalry network from Baseline prediction, and (c) Rebuilt rivalry network using Naive Bayes classifier.

rival gang territories while moving more freely inside non-rival territories. The activity matrix is further normalized to contain fraction of activity points instead of absolute counts: For $i, j = 1, 2, .., n$, we have $A_{i,j} = \frac{A_{i,j}}{\sum_j A_{i,j}}$

Given the small area of Hollenbeck, the calculated activity matrix is very sparse. After ignoring gang territories with only one user, clipping links with very few activity points, and balancing the number of rivalry links and non-rivalry links, we came up with a subgraph G' of G with 54 links (27 rival and 27 non-rival). With each of these links as an instance and features calculated using the activity matrix, we train a Naive Bayes classifier.

B. Link Features

For each link between two gangs g_1 and g_2 we calculate the following features:

- *Distance-based features*: Centroid distance (distance between the centroids of the two gangs g_1 and g_2's territory); Closest distance between two gang territories; Maximum territory span among the two gangs (territory span being the distance between the two furthest points in a gang's territory); Squared maximum territory span; Absolute difference between territory spans.
- *Activity-based features*: Total activity ($A_{1,2} + A_{2,1}$); Average activity ($\frac{A_{1,2}+A_{2,1}}{2}$).
- *Network-density features*: Sum of incoming density of nodes g_1 and g_2 in graph G; Absolute difference of incoming densities; Sum of outgoing densities; Absolute difference of outgoing densities. Here incoming density and outgoing density are calculated as,

$$InDensity(\text{node}) = \frac{\text{No. of incoming edges}}{\text{No. of possible edges}} \quad (3)$$

$$OutDensity(\text{node}) = \frac{\text{No. of outgoing edges}}{\text{No. of possible edges}} \quad (4)$$

- *Entropy-based features*: Sum of entropy of incoming edges to g_1 and g_2; Sum of entropy of outgoing edges from g_1 and g_2; The in-entropy and out-entropy for a gang g_i is given by:

$$H^{in}_{g_i} = -\sum_{j,j\neq i} ln(A_{j,i})A_{j,i} \quad (5)$$

$$H^{out}_{g_i} = -\sum_{i,i\neq j} ln(A_{i,j})A_{i,j} \quad (6)$$

- *Cross-entropy-based features*: Cross entropy of incoming edges to g_1 and g_2; Cross entropy of outgoing edges to g_1 and g_2. The in-cross entropy and out-cross entropy for gangs g_i and g_j is given by:

$$H^{in}_{g_i,g_j} = -\sum_{k,k\neq i,k\neq j} ln(A_{k,i})A_{k,j} \quad (7)$$

$$H^{out}_{g_i,g_j} = -\sum_{k,k\neq i,k\neq j} ln(A_{i,k})A_{j,k} \quad (8)$$

C. Prediction Results

The baseline method against which we would be comparing our results is derived from the intuitive reasoning that a street gang is often in conflict with other territories which share a

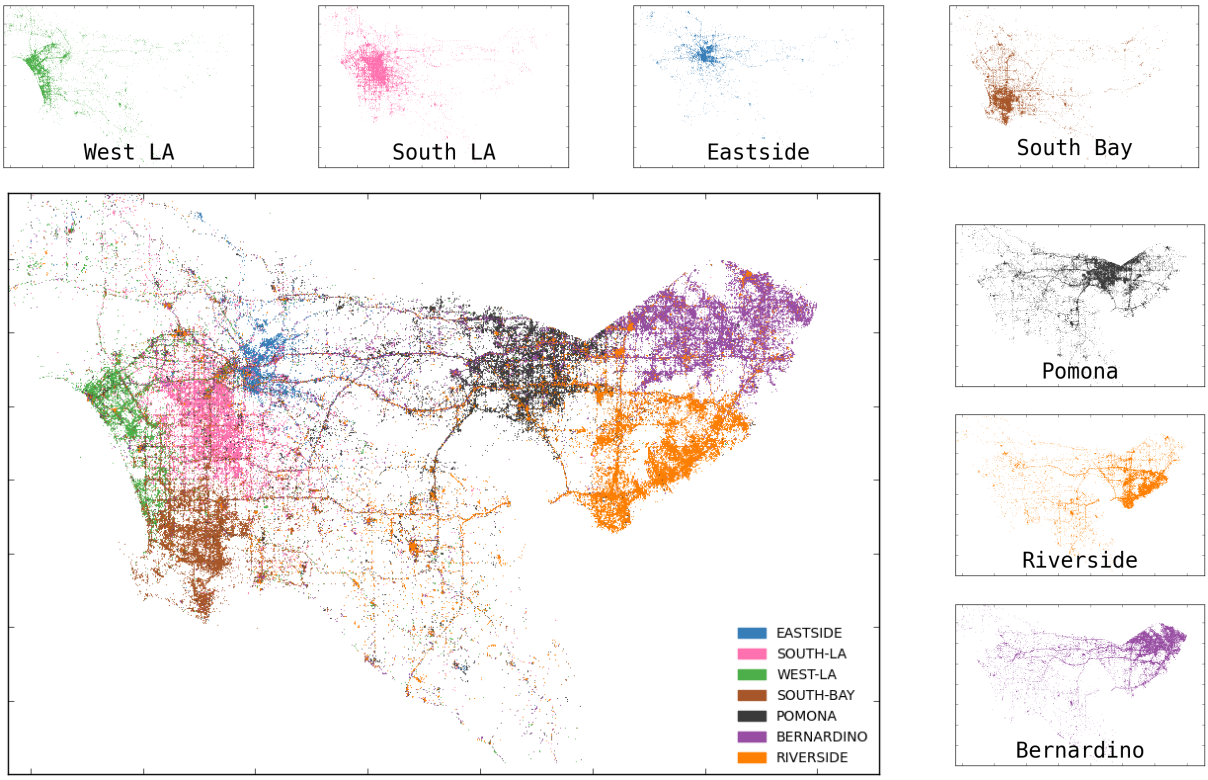


Fig. 4: Figure shows scatter plots of all tweeting activity by users belonging to seven regions in Los Angeles- Eastside, South LA, West LA, South Bay, Pomona, San Bernardino and Riverside.

common geographic boundary with it, so that it can expand its territorial area. The baseline method marks each such pair of territories as rivals when they share a border, and all remaining pairs as non-rivals. Using the features discussed above, we were able to achieve a leave-one-out training accuracy of 87% over all rinks. The baseline performs at 78% so we have a 9% improvement. Looking at the rivalry network that was rebuilt using both these methods, i.e. considering only rival links, which are important to discover, the baseline system was able to rebuild only 60% of the original, while the classifier model could rebuild 89% of the original rivalry network (Figure 3), which is a 29% improvement.

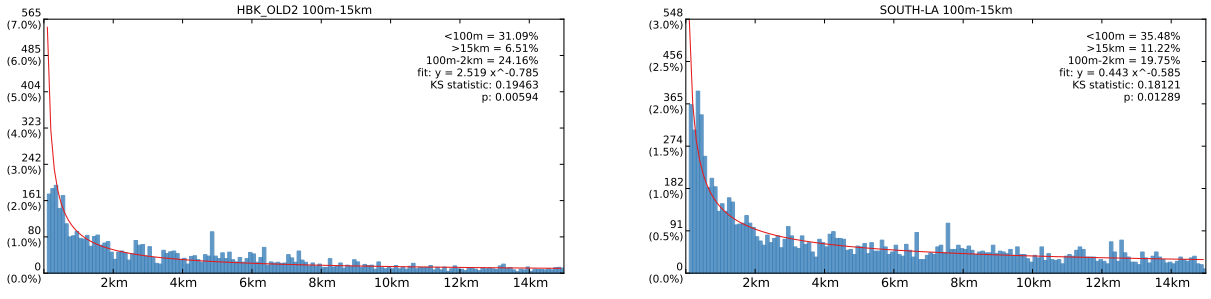
A question that arises now is how such a model could be useful. As pointed out in [12], the street gangs territories are dynamic in nature, and can be attributed to multiple factors like relocation, admission and dismissal of gang patterns. The data on which the rivalry network is based on may be old, and the changing rivalries might not be accounted for immediately. Furthermore, since most of the street gang's activity are criminal in nature, any unreported incident will skew the data collected by law enforcement. In such situations, using a previously validated model and real-time data from location sharing services would help keep in pace with changing rivalries and hence, effective policing.

V. Understanding Patterns in Mobility

Having established how Twitter data can be used to model street gang relationships, their effect on mobility of people living inside them and in turn predict the rivalry network between them, we now expand the geographic resolution and try to see what insights the data might provide on a larger canvas. We pick seven regions in Los Angeles- Eastside, Pomona, Riverside, San Bernardino, South Bay, South LA and West LA, and assign users to each of these regions. Similar to the previous experiment, a user belongs to a particular region if his/her home location lies within the geographic boundary of the region. Figure 4 shows a scatter plots of all tweeting activity by users belonging to each of these seven regions. Apart from roughly shading out the geographic boundary of the city, this plot also reveals something more interesting- the interpolation of all tweeting activity highlights all major roadways in the cities, even pointing up smaller streets and intersections. This observation alludes the fact that people apparently tweet a lot while driving.

Figures 4 show another interesting difference between the patterns and extent of commute followed by people belonging the two regions: West LA and Pomona. Users from West LA mostly stay within the region, and barely travel east beyond mid city. However, in case of the region Pomona, there is a larger activity far west, showing that people belonging to this region would travel about 30 miles to reach all the way to areas like Downtown Los Angeles, and Long Beach. Segregating the seven regions we picked before into two groups, Eastside, South Bay and South LA shows similar behavior to West LA, whereas the regions far east (San Bernardino, Pomona and Riverside) share similar properties.

An established characteristic of human mobility patterns is



(a) Eastside

(b) South LA

(c) West LA

(d) South Bay

Fig. 5: Figure shows the distance from home histograms for each tweeting activity belonging to four regions in Los Angeles- (a) Eastside, (b) South LA, (c) West LA, (d) South Bay.

the approximation of trajectories of travel using Levy flight and random walk models. Often used to approximate movement and migration patterns in birds and animals, the Levy flight model states that the trajectory of motion follows a sequence of random steps, where the step size belongs to a power law distribution (Equation 9 shows the probability distribution function for the power law). In other words, there are a large number of short hops and fewer long hops. In [6] and [1], they approximated truncated power law distributions for distance and duration of steps in a trajectory.

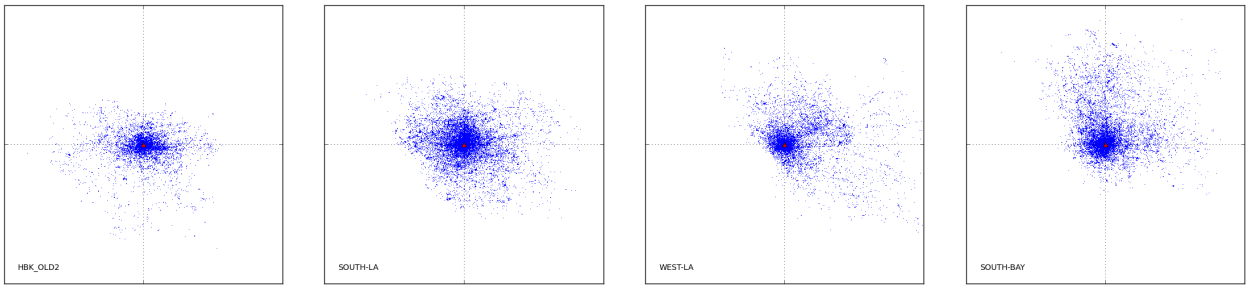
$$f(x) = Cx^{\alpha} \tag{9}$$

We modify their approach by gauging distance from home for each activity point, instead of distance between consecutive activity points. Once again, we pick four regions (Eastside, South Bay, South LA and West LA), along with the users belonging to each of them. Figure 5 shows the distance from home histograms, over the range 100m to 15km, for each tweeting activity belonging to these four regions. Activity points within 100m from home location were removed as such small shifts may be due to GPS noise even when the user is stationary. The distribution beyond 15km dwindles and becomes sparse, hence removed as well. Similar to the observations in [6] and [1], the displacement from home function also shows heavy tailed characteristics, i.e., the distance from home follows a power law distribution. We verify this using a two-sample Kolmogorov-Smirnov test, where the null hypothesis is that the samples are drawn from the continuous power law distribution. In each case, the null hypothesis was accepted with significance ($p < 0.05$), hence verifying correctness of the least square estimates for power law.

Although we can model distance traveled by a user from his/her home using a power law function, this isn't enough to model user mobility on a two dimensional geographic canvas. The second element that needs to be modeled along with distance, is the direction of travel. Figure 6 shows scatter plots of all activity points by users in each of the four picked regions. These plots were generated by considering separate layers for each user and his/her activity points, and then shifting these layers so that the home location for all users overlap at a single point- the center coordinate of the plot.

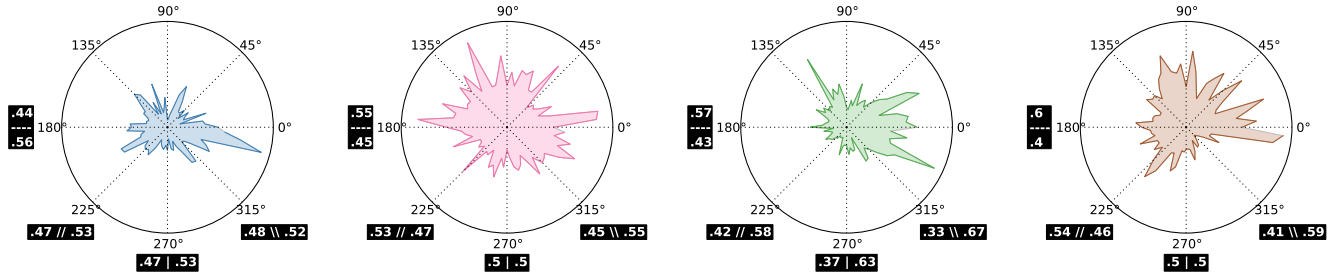
In [12], while studying formation and evolution of rivalry networks in gang territories, they found that freeways often act as natural barriers limiting any activity permeating through it. The same holds true for physical barriers like mountains and coastlines. Looking at Figure 6(a) and 6(b), we see that the direction of displacement is almost uniform in all direction. Both Eastside and South LA are centrally located within the city and do not have physical barriers on any side to obstruct the direction of travel. However, West LA has the Pacific coast running its length onto its west, and South Bay is similarly bounded on the west and south. The effect of these barriers is visible as a clear attenuation on the scatter plots 6(c) and 6(d).

Trimming the obvious high concentration of tweets right around home, we were able to generate the distribution for direction of activity in each of these regions. Figure 7 shows the probability density function (PDF) for direction of displacement as a radial plot for each of the four regions. The figures also shows the ratio in which the four diagonals, N-S, NE-SW, E-W and SE-NW, split the distribution on either side of them. For Eastside and South LA (Figure 7(a) and 7(b)) the PDF function is uniform and almost evenly split by all of



(a) Eastside (b) South LA (c) West LA (d) South Bay

Fig. 6: Figure shows scatter plots for tweeting activity belonging to each user (with home location shifted to center of plot) of four regions in Los Angeles- (a) Eastside, (b) South LA, (c) West LA, (d) South Bay.



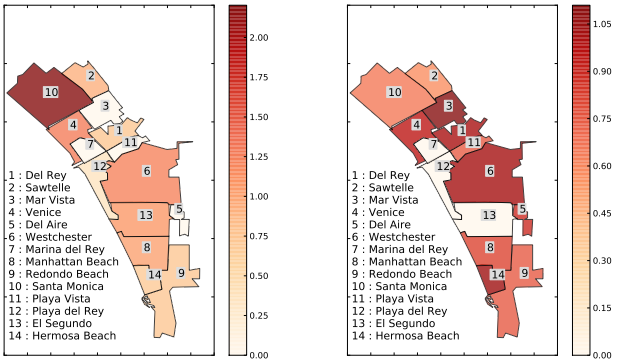
(a) Eastside (b) South LA (c) West LA (d) South Bay

Fig. 7: Figure shows the direction of displacement from home Probability Distribution function for each tweeting activity belonging to four regions in Los Angeles- (a) Eastside, (b) South LA, (c) West LA, (d) South Bay.

the four diagonals. However, West LA (Figure 7(c)) shows a very strong skew across the SE-NW and N-S diagonals, which mirrors the coastline on the west. For South Bay (Figure 7(d)) the skew is not as apparent as that of West LA, but there does exist a skew across the SE-NW diagonal and the E-W diagonal, each of them parallel to the coastline on west and south respectively. This show that directional distributions might be a valuable source to extract regional constraints of neighborhoods. In the case of gangs there were social deterrents and in this case, we have geographic deterrents.

Apart from modeling and distinguishing mobility patterns of users region-wise, we can also account for social interactions and travel patterns of people living in their constituent neighborhoods by measuring visits between them. We pick the West LA region as a case study, and start by assigning users to each neighborhood similar to that explained in Section IV. We then calculate an activity matrix using Equation 2, and for each neighborhood calculate entropy-in and entropy-out as described in Equation 5 and 6. Figure 8 a gradient map of the West LA neighborhoods ranked by entropy-in and entropy-out features.

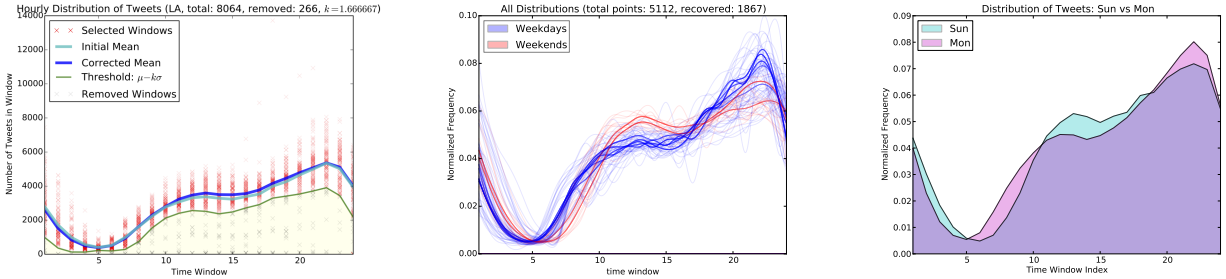
A high entropy on incoming visits signifies uniform visit pattern by people belonging to all other neighborhoods. Santa Monica has the highest entropy-in, which is understandable given it is a popular beach location and draws visitors from all areas. This in some degree reflects the local truth that it is a hub neighborhood of West LA. Venice follows Santa Monica closely, which also is another popular beach location.



(a) Entropy in (b) Entropy out

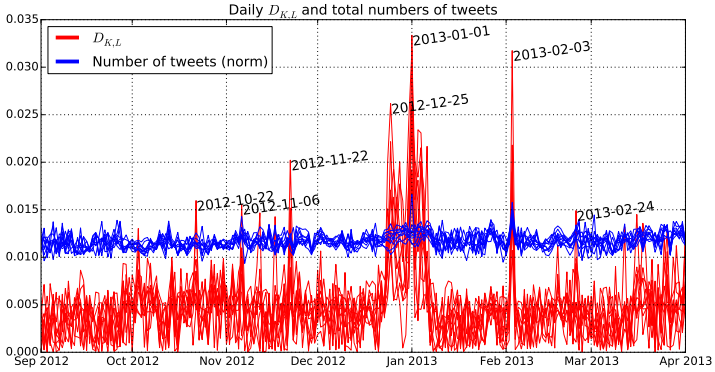
Fig. 8: Figure shows color map of neighborhoods in West LA ranking features (a) Entropy In and (b) Entropy Out calculated on visits between all neighborhoods in the region.

Westchester, which houses the Los Angeles International airport, also shows high entropy. On the other hand, Mar Vista is a rather quiet residential area, and probably isn't visited by many people, hence the extremely low entropy-in and also the high entropy-out as they tend to go to all the surrounding areas. Westchester also has a high entropy-out, which is explained by it containing an university, which can be clearly identified by the cluster of home locations near it, that has students that tend to travel to all the surrounding areas.



(a) Mean tweeting pattern (b) Seven days of week (c) Sunday vs Monday

Fig. 9: Figure shows tweeting rate for a day over all of Los Angeles.



Dates	Description
Oct 22, 2012	3rd Presidential debate
Nov 06, 2012	Elections
Nov 22, 2012	Thanksgiving
Dec 25, 2012	Christmas
Jan 01, 2013	New Year
Feb 03, 2013	Super Bowl
Feb 24, 2013	Oscar

Fig. 10: Figure shows events identified using the KL divergence method during the period from Sep'12 to Apr'13.

VI. Event Detection using KL Divergence

Using our data to model user mobility gives us insights of its utility in the spatial dimension. The complementary temporal dimension of tweeting patterns is equally interesting. Looking at the tweets over a span of months, we noticed that the frequency of tweets on the time axis follows a regular distinctive shape. While this characteristic shape remains similar through weekdays, the weekend tweeting patterns is slightly different. We will now model the daily distribution of tweet frequency over the regions of Los Angeles and use it to identify events by detecting any skew in the tweeting pattern on a particular day.

A. The Approach

The first step of our approach is to estimate the average distribution of tweets for a given day of week over a region. Occasionally, our data collection process would be turned off for some time period, creating gaps in tweet frequency over the time axis. These gaps need not necessarily be zero-frequency values, but can simply be a skewed drop from the regular pattern. To ensure that these gaps do not distort our estimate of mean tweet frequency, we eliminate these skewed time periods by using a minimum frequency threshold equal to $\mu - k * \delta$, where μ is the mean for that time period and δ the standard deviation. Figure 9(a) shows the threshold and how including the threshold shifts the estimate of mean distribution slightly. Figure 9(b) shows the average distribution of tweets on a time window of 24 hours for each of the 7 days of the week. Figure 9(c) shows the difference in the tweeting pattern on a weekend and an weekday more clearly.

For a particular day, we can now compare the distribution of tweets to the mean tweeting distribution for that day of week calculated over a region. For example, the distribution of a specific Friday in Riverside will be compared against the mean distribution model for a Friday in Riverside. For comparison, we use the Kullback-Leibler (KL) divergence for two distributions P and Q given by the following expression:

$$D_{KL}(P||Q) = \sum_i ln(\frac{P(i)}{Q(i)})P(i) \tag{10}$$

The activity in a particular region on a day can be marked as being "unusual" if there is a spike in KL divergence on that day. Figure 10 shows the KL divergence plots for all regions over the time period of September 2012 to April 2013. It also shows the total number of tweets for those days, normalized to fit on the graph. We can see that the KL divergence methods shows spikes that are not visible by just looking at the total tweets. Most of the prominent spikes correspond to some popular holiday and big events, when the spikes are coherent in all regions. Apart from the prominent large spikes, there are a number of smaller spikes, and spikes that occur only in a single region which represent smaller events, breaking news or some local trend. The adjoining table in Figure 10 accounts

for the events on some of these dates. This shows that the KL divergence approach can discover outliers that are invisible to simple approaches like tweet counts. The key is doing this without looking at the text of the tweets, only the temporal patterns of when they were sent.

VII. Conclusion

Adhering to the primary objective of our study, we were successfully able to use ubiquitous geo-location information acquired from social media, to model human behavior and mobility patterns. In doing so, we also encountered a number of interesting insights. We collected strictly geo-tagged tweets within the geographic boundary of the city of Los Angeles using Twitter's Streaming API, and used DBSCAN clustering algorithm to compute home locations for unique users all over the city. We then modeled users as belonging to different regions and neighborhoods, and used their tweet locations to attribute their activity within particular geographical spaces.

Our first experiment was regarding street gangs of Hollenback policing district in Eastside. Mapping users to street gang territories and their tweeting activity as visits to all other gangs territories, we trained a machine learning classifier to tell apart links between gangs as being of rival or non-rival nature. Using this method we were able to reconstruct 89% of the actual rivalry network, which beats the baseline results by around 30%. Looking at a larger geographic canvas, we noticed that interpolating all tweeting activity on a plot shaded out all major roads and intersections throughout the city, a result hinting at the high volume of tweet while driving. Also, the distance from home distribution of all tweeting activity showed heavy-tailed nature following a power-law distribution, which is similar to previously established nature of distance and time between hops in a trajectory. Another interesting observation was that the direction of displacement of users is random in direction and uniform in order, only to be skewed by the presence of physical barriers like the freeways, coastline or mountains. We were also able to establish variations in travel patters of users belonging to the different regions, and how a simple entropy-based metric can be used to find popular destination among neighborhoods.

Another experiment we undertook was to identify events by comparing the tweet frequency distribution of a day to the mean distribution of that day of week. The tweet frequency over 24 hours of a particular day of week follows very characteristic curve. Comparing the distribution of a day to the mean distribution for that day in a region using KL divergence, we were able to identify popular holidays, major US events and few smaller events spread across the city. This method can find outliers that one may not find by just looking at tweet counts. The work here indicates the promise of using spatiotemporal social media as a new source to discover interesting patterns about human behavior and also to characterize various geographic phenomena. The ever growing geo-tagging of social media through other sources, such as photoblogging, and the integration of multiple social media sources could lead to better ways of understanding our cities by discovering the innate patterns of life within them.

Acknowledgements. This material is based upon work supported in part by the AFOSR under Award No. FA9550-10-1-0569 (Bora, Zaytsev, and Chang) and the ARO under MURI award No. W911NF-11-1-0332 (Maheswaran).

References

[1] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[2] J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh, "The livehoods project: Utilizing social media to understand the dynamics of a city," *ICWSM'12*, 2012.

[3] M. Short, M. DOrsogna, P. Brantingham, and G. Tita, "Measuring and modeling repeat and near-repeat burglary effects," *Journal of Quantitative Criminology*, vol. 25, no. 3, pp. 325–339, 2009.

[4] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, "Self-exciting point process modeling of crime," *Journal of the American Statistical Association*, vol. 106, no. 493, 2011.

[5] A.-L. Barabasi, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, no. 7039, pp. 207–211, 2005.

[6] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.

[7] A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, and A.-L. Barabási, "Modeling bursts and heavy tails in human dynamics," *Physical Review E*, vol. 73, no. 3, p. 036127, 2006.

[8] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the levy-walk nature of human mobility," *IEEE/ACM Transactions on Networking (TON)*, vol. 19, no. 3, pp. 630–643, 2011.

[9] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[10] R. Ahn, P. Elliott, and K. Luh, "Social network clustering: An analysis of gang networks," 2011.

[11] Y. van Gennip, B. Hunter, R. Ahn, P. Elliott, K. Luh, M. Halvorson, S. Reid, M. Valasik, J. Wo, G. E. Tita *et al.*, "Community detection using spectral clustering on sparse geosocial data," *SIAM Journal on Applied Mathematics*, vol. 73, no. 1, pp. 67–83, 2013.

[12] P. J. Brantingham, G. E. Tita, M. B. Short, and S. E. Reid, "The ecology of gang territorial boundaries," *Criminology*, vol. 50, no. 3, pp. 851–885, 2012.

[13] L. M. Smith, A. L. Bertozzi, P. J. Brantingham, G. E. Tita, and M. Valasik, "Adaptation of an animal territory model to street gang spatial patterns in los angeles," 2012.

[14] G. Tita, K. J. Riley, K. Jack, G. Ridgeway, C. Grammich, A. F. Abrahamse, P. Greenwood, and R. Corporation, "Reducing gun violence: Results from an intervention in east los angeles," 2003.

[15] G. Tita, J. Riley, G. Ridgeway, and C. Grammich, "Unruly turf: The role of interagency collaborations in reducing gun violence," *Rand Review*, 2003.

[16] K. Louie, M. Masaki, and M. Allenby, "A point process model for simulating gang-on-gang violence," 2010.

[17] M. Egesdal, C. Fathauer, K. Louie, J. Neuman, G. Mohler, and E. Lewis, "Statistical and stochastic modeling of gang rivalries in los angeles," *SIAM Undergraduate Research Online*, vol. 3, pp. 72–94, 2010.

[18] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 851–860.

[19] D. Chakrabarti and K. Punera, "Event summarization using tweets," in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011, pp. 66–73.

[20] R. Li, K. H. Lei, R. Khadiwala, and K.-C. Chang, "Tedas: a twitter-based event detection and analysis system," in *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*. IEEE, 2012, pp. 1273–1276.

[21] J. Lanagan and A. F. Smeaton, "Using twitter to detect and tag important events in live sports," in *Proceedings of AAAI*, 2011.

[22] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." Kdd, 1996.