Crime Hotspot Tracking and Geospatial Analysis in Merseyside, UK

Michael W. Milo (mmilo@vt.edu)†, Sarah C. Richards (scrich@vt.edu)‡, Parang Saraf (parang@vt.edu)*, Lonesome Malambo (mmoonga@vt.edu)§, and Tirtha Bhattacharjee (tirtha23@vt.edu)*

†Department of Computer Science ‡Department of Mechanical Engineering ^Department of Statistics §Department of Biological Systems Engineering

Abstract—Crime prediction is a topic of significant research across the fields of criminology, data mining, city planning, law enforcement, and political science. Crime patterns exist on a spatial level; these patterns can be grouped geographically by physical location, and analyzed contextually based on the region in which crime occurs. This paper proposes a mechanism to parameterize street-level crime, localize crime hotspots, identify correlations between spatiotemporal crime patterns and social trends, and analyze the resulting data for the purposes of knowledge discovery and anomaly detection. The subject of this study is the county of Merseyside in the United Kingdom, over a span of 21 months beginning in December 2010 (monthly) through August 2012. Several types of crime are analyzed in this dataset, including Burglary and Antisocial Behavior. Through this analysis, several interesting findings are drawn about crime in Merseyside, including: hotspots with steadily increasing crime levels, hotspots with unstable crime levels, synchronous changes in crime trends throughout Merseyside as a whole, individual months in which certain hotspots behaved anomalously, and a strong correlation between crime hotspot locations and borough / postal code locations. We believe that this type of statistical and correlative analysis of crime patterns will help law enforcement agencies predict criminal activity, allocate resources, and promote community awareness to reduce overall crime rates.

I. INTRODUCTION

It is well known that individual crimes are not unique random events but rather share a number of common characteristics [1]. The spatial distribution of certain crime types within a given area is of great interest to law enforcement as well as other government agencies and the understanding of this distribution is crucial to effective policing. Recent research in the area of crime clustering has shifted focus from large administrative areas to much smaller geographic regions [2]. These new “micro-units” [2] of interest can be neighborhood- or even street-level areas and can be analyzed to visualize small-scale crime hotspots that may indicate an increased level of crime within the area. Research also suggests that focusing presence to these crime hotspots can lead to significant crime prevention gains [3].

Data containing point-locations of individual crimes is necessary for such small-scale analysis and has been increasingly more available over the past several years. Many data sources are lacking complete datasets or require significant cleaning before use. In January 2011, the United Kingdom’s Home Office launched http://www.police.uk, a website containing individual crime locations for the entire nation [4], allowing for spatiotemporal analyses of crime within the country. In this study, we focus on the county of Merseyside, a small, metropolitan county in Northwest England (Figure 1). Merseyside contains a mixture of high-density urban areas associated with Liverpool, suburbs, semi-rural, and rural areas [5]. This geographically diverse county provides unique environments for the potential of crime hotspots. Data for total crime, burglary, and antisocial behavior are analyzed and clusters of each crime type for each month are generated. A contextual analysis of these clusters over time will not only provide insight into the “why” of criminal activity, but provide law enforcement and local government agencies a visualization of crime hotspots and “lowspots”, allowing for intelligent resource allocation dedicated to preventing and reducing crime within the clusters and Merseyside as a whole.
II. RELATED WORK

Clustering in general provides valuable information about spatially-correlated data in crime analysis. However, the existence of a wide variety of approaches to clustering makes it difficult to label just one as superior. Previous literature has demonstrated success using the K-means algorithm, Mixture Models, and hierarchical grouping methods. Other methods have succeeded in localizing crime to hotspots using local crime point-densities [1], tracking small-scale vectors of crime rates on a street-by-street basis [6], and assigning crime densities to a geospatial grid for preventative targeting [7].

While crime clustering has been an active research topic, many hot-spot analysis techniques have often treated the spatial and temporal aspects of crime as distinct entities, thus ignoring the necessary interaction of space and time to produce criminal opportunities. However, many studies have stated that recognizing the differences in the spatiotemporal signatures of crimes, and how they vary, can deepen our understanding of criminogenic processes. To demonstrate the importance of the space-time factor to crime hotspot analysis, Grubesic and Mack [8] utilized a two spatiotemporal methods, the Knox test and the Jacquez k-nearest neighbor test, to evaluate the space-time footprints of burglaries, assaults and robberies in an urban environment based on a comprehensive database of crime events for seventeen neighborhoods in Cincinnati, Ohio. Knox test evaluates spatial-temporal interaction using specified distance and time thresholds while the Jacquez uses N nearest neighbors in both space and time. Results of the study suggested that robbery, burglary and assault have dramatically different spatiotemporal signatures [8].

Some studies have looked at the response of crime patterns to police interventions. Wyant et al. [9] evaluated the association of firearm arrests with subsequent shooting for the city of Philadelphia. The data used in this study comprised shootings, including criminal homicides, robberies, and aggravated assaults by firearm, as well as Violations of the Uniform Firearms Act (VUFAs) occurring within the city from January 1, 2004 to December 31, 2007. At a spatial distance of 400 feet and a temporal resolution of 2 days, a modified Knox close-pair method was used to test whether there was clustering of events in both space and time. Results showed that following a firearm arrest there was a significant decrease in shootings by as much as 28-47% up to a couple of blocks away but were short-lived.

Mohler et al. [10] adapted self-exciting point processes, normally applied by seismologists to study earthquakes, to model spatial-temporal clusters of crime in Los Angeles. Crime patterns were modeled as a space-time Poisson process of background events, each triggering a sequence of aftershocks analogous to those in seismology. Their findings show that this approach provides a more complete picture of the statistical nature of crime and has important implications for crime prediction and prevention.

Geographic information systems are also playing a big role in crime analysis because of their capability to handle spatial data. Cusimano et al. [11] examined the spatiotemporal dimension of violent crime in Toronto, Canada. An ambulance dataset of 4,587 ambulance dispatches and 10,693 emergency room admissions for violent injury occurrences among adults (aged 18-64) in city during 2002 and 2004 were used in the study. Kernel density and choropleth maps for 24-hour periods and four-hour daily time periods were created and compared with location of ambulance dispatches and patient residences with local land use and socioeconomic characteristics. A multivariate regression was used to control for confounding factors. Results showed that locations of violent injury and the residence locations of those injured were both closely related to each other and clearly clustered in certain parts of the city characterized by high numbers of bars, social housing units, and homeless shelters, as well as lower household incomes.

Contextual analysis and social networks have also been leveraged to infer information about the structure of criminal networks. However, the scope of these analyses has been limited only to social data, omitting geospatial crime rate data [12]. The National Consortium for the Study of Terrorism and Responses to Terrorism uses several web resources and forums to collate and extract information pertaining to terrorist groups. They identify several event-level and actor-level information to make operational decisions and design counter-terrorism strategies [13].

III. PROPOSED APPROACHES

We propose a twofold method for crime analysis and prediction. In this study, we first analyze crime in a spatiotemporal framework to parameterize crime hotspot locations, shapes, point assignments, and other factors of interest. After a comprehensive analysis of the spatiotemporal clusters yields actionable information, these analyses are validated against existing known crime data using borough and postal code boundaries as contextual references.

In order to predict crime in both spatial- and temporal-aspects of street-level crime, analysis of crime data should incorporate multiple predictive elements in order to classify crime trend aspects. Such a system should:

- Localize distinct types of crimes to geographic regions
- Assess likelihoods for which types of crimes occur within these regions
- Track regional trajectories and predict future crime

While clustering is a widely-used tool to accomplish the goal of localization and parameterization, most clustering methods fall short of automatically determining an appropriate number of clusters to apply to the data. The problem of discovering how many clusters exist in a particular dataset
is nontrivial and accounts for a large volume of existing research. Entropy estimators yield a rough estimation of the tipping point at which additional clusters yield no significant analysis advantage. The nonparametric approach of measuring the distortion of the dataset – that is, the average distance between each data point and its closest cluster center – has proven a more efficient and effective method for estimating $K$ [14].

Further, educational, economical, demographic, land use and health data will also be used to identify the potential reasons for seeing particular crime patterns for a location. As an example, Southport, UK shows the maximum concentration of antisocial behavior in the whole of Merseyside. Using the economical and land-use data to identify causes for this trend (Southport is a popular seaside tourist destination), it is possible Southport has underlying factors which cause it to exhibit abnormally high antisocial behavior. We will try to identify similar examples for the several crime trends and patterns that we observe for the whole of Merseyside, UK using both clustering analysis results and contextual inference.

Our proposed method will develop data-driven clustering analysis, applied to various types of geo-tagged data (crime data from www.police.uk/data, regional demographics, etc.) and attempt to estimate a value for $K$ clusters in the data. By parameterizing these clusters as distributions, we will be able to assign a confidence rating to crime type likelihood in specific areas through Expectation Maximization and classification analysis. The result of this analysis will be twofold:

1) Crime rate localization and parameterization

2) Geospatial social, economic, demographic, etc., cross-validation

By combining these results, our work seeks to find some measure of correlation between crime categories and spatiotemporal crime trends. These correlative measures may be found through any one of several techniques: spatial autocorrelation, distribution matching, and other correlative metrics such as the Pearson Correlation Coefficient. Finally, temporal analysis demonstrates trends in the clustering results (number and shape of clusters, individual cluster movement, etc.) to facilitate future crime prediction, as well as pattern recognition in social media (local tweet counts, keyword frequency, etc.) to facilitate anomaly detection.

IV. SYSTEM DESIGN

The structure of our analysis system contains three key stages explained in detail in the following sections. The initial stages of this study were focused primarily on data gathering and selection, choosing which data were appropriate for study, cleanup, etc. The core of this paper focuses on the analysis of the data, knowledge discovery, contextual inference, and case studies. Finally, this information is presented to a user through a web-based User Interface (UI), which allows a user to analyze large portions of data at once, and draw conclusions about complex data relationships.

The workflow of this study is shown visually in Figure 2.

A. Data Selection

A wide variety of crime data sources are available for public use. Prior to the analysis portion of the study, it was necessary to first determine which of these many data sources would be most appropriate for street-level crime trend analysis, with the possibility of incorporating social media or contextual information at a later time.

Among these available data sources were:

- **FBI Uniform Crime Reporting (UCR) Database:** The FBI’s UCR database acts as a yearly repository of crime data in the United States, as reported by individual police precincts in a specified format.

  While this dataset is an extensive model for crime in the United States, it was not selected for this study because the data are not granular enough, and too broadly classified to draw street level trends.

- **Crime Data from Mark Everline, 2010:** Mark Everline (et al.) has provided a pre-pased database of crime from the Washington D.C. area from 2009 through 2010.

  This dataset was not selected because data parsing is irregular and contains many missing elements; it would require data omission or cleanup to fit an analysis model.

- **Crime data from Brewer et al., 2011:** This antiterrorism dataset was not selected because of its vast geographical range; it would be impossible to classify
street-level crime trends when crime types vary by country.

Additionally, the span of this data (< 12 months) is insufficient to draw adequate time trend conclusions, which are crucial to this study.

The dataset chosen for this study was compiled by the United Kingdom Police and Home Office, and contains street-level crime data segmented by county. Of the more than 80 counties available in this dataset, our work focuses on information from Merseyside, U.K. (a subject of previous crime trend analyses). This dataset spans a period of 21 months beginning December 2010 and running through August 2012. An additional benefit of this dataset is the incorporation of August 2011, the month in which the London Riots occurred (the effects of which were felt throughout the U.K., and are incorporated into case studies in this paper). Finally, this dataset has been the subject of other studies to analyze local and street-level crime in other localities throughout the UK. One example of which is shown in Figure 3, which is another web-based application found at oobrien.com/vis/crime. This web-based app only shows the local density tracking of various crime types (e.g., Burglary) in the greater London area; our study seeks to parameterize such local densities and categorize them as crime “hotspots”.

This dataset is stored in parsed Comma-Separated Value (CSV) format, with street-level crime incidents organized by county and month, and stored as Easting and Northing coordinates (British National Grid). The data are publicly available at police.uk/data, and are labeled according to crime type (an example repository page is shown in Figure ??

Because of the variability of crime type nomenclature – some months group “drug” and “arson” crime with “other crime”, while others delineate between all types – only three crime types were selected for study in this paper:

- **Burglary**: theft of goods or materials from a property, often through the act of breaking and entering or forced entry. This is not to be confused with “robbery”, which requires some threat or implication of harm.

- **Antisocial Behavior**: behavior that shows a lack of consideration for others with intent to cause damage to society. This is similar (but not identical) to the American classification of “disorderly conduct”.

- **All Crime**: all crime types aggregated into one single stream, stripped of all label data and indexed by month. This global category yields information about the general crime in a region, and acts as the basis for identifying clusters within that region.

### B. Data Analysis

The proposed data flow for this study involves the clustering of spatial crime data to identify hotspots, parameterization of hotspots, and tracking of spatiotemporal trends in hotspots. Using these geospatial regions-of-interest, contextual and social media information can be indexed to infer additional knowledge about the causes of crime. The data flow for this study takes several key steps:

1. **Classify individual geospatial crime events by date and crime type**
2. **Estimate spatial clustering characteristics (i.e., number of clusters, $K$) of crime events**
   1. Use Sum of Squared Error (SSE) for all values of $K$
   2. Select a value of $K$ at which SSE becomes stable (no further information gain takes place)
3. **Using experimental values for k and centroid locations**
   1. Fit the spatial crime map to a mixture $K$ model using Expectation Maximization
   2. Obtain a parameterized model for crime hotspot locations, shapes, etc.
   3. Track these parameters over time, and predict how crime is changing
4) Using spatial boundaries, extract social and contextual information from specific areas
   a) Use the contextual information to determine causes of crime (e.g., bars, tourism)
5) Feed all of this data into a user interface (See Section IV-C)
   a) Allow the user to select layers of information to view at the same time
   b) Provide scalable maps of study regions at street, block, town, and county levels

Cluster centroid numbers and locations are estimated in order to track crime from month to month. This approach makes the key assumptions that:

- Crime hotspots are generally well-correlated between months
- Crime hotspots are neither appearing nor disappearing
- Hotspot shape parameters are generally stable over time
- A crime “hotspot”, should have at least one crime associated with it for any given month

Figure 5 shows an example map of Merseyside, UK and 6 arbitrarily-chosen points to act as sample hotspots. This data would be obtained from the application of clustering algorithms such as K-means, Gaussian mixture models, and nearest-neighbor networks. Measures of skewness and information entropy provide an estimator for the true number of clusters, \( K \). This is a nontrivial problem, and has been the subject of significant research in the fields of Machine Learning and Data mining. Our proposed approach utilizes a computationally efficient solution for estimating \( K \) based on the Sum of Squared Error of the K-means algorithm result, discussed further in Section V-A.

The measured chaos, or entropy, of all clusters is monotonically decreasing on the range \( K = [1; N] \) where \( N \) is the number of instances in the dataset. As this calculation decreases, its second derivative reaches local (and global) maxima; these “elbow points” serve as estimators for the appropriate value for \( K \). Recent literature has also implemented skewness as an estimator for \( K \). After fitting to a mixture model, the Mahalanobis Distance compares each cluster to its expected covariance through matrix rotation.

The result of these metrics should yield an appropriate value (and locations) for the number of crime hotspots. Using these cluster centroids, each month is processed and fitted to a mixture model using Expectation Maximization (EM) algorithms. Gaussian Mixture Models (GMM) are well-suited to describing spatial point data, and provide shape information through the covariance matrix (calculated during the EM process). Tracking the changes in mean location, shape data, cluster size, number of points contained, and point density yield predictive trend statistics over a period of multiple months. Figure 6 shows one such example progression of 6 hotspots tracked over the course of 3 months.

1) Clustering with K-Means: The application of basic clustering algorithms such as K-means yields information about the number of clusters in the dataset, and their locations. The K-means algorithm is a form of Expectation Maximization, in which the Expectation step assigns each point of the dataset to its nearest cluster center, and the Maximization step calculates new centroids based on the distribution of point assignments. This process repeats for a user-specified number of iterations, or until a specified tolerance is reached. The K-means algorithm may not reach a globally optimal solution, and instead become stuck in local minima. As the algorithm is fairly numerically well-behaved and computationally inexpensive, it is common to perform multiple iterations of the algorithm to obtain better estimates of the true data clustering schema. An example K-means algorithm is found in Figure 7.

Selection of various values of \( K \) yields very distinct cluster maps, as shown in Figure 8. Even at a preliminary stage, plotting these clusters against geospatial map data yields additional information, an example of which is shown in Figure 9.

2) Expectation Maximization: The Expectation Maximization algorithm fits point index data to a Gaussian
X = input data;
K = input # of means;
// initial mean locations & assignments:
initialize MUs, XK;
// loop counter & max loops:
initialize l, maxL;
// centroid motion & tolerance:
initialize e, tol;
while e < tol && n < maxN {
    // Expectation step
    for i=1, i<=length(X), i++ {
        dMUs = X(i)-MUs;
        [V, I] = min(dMUs);
        Xk(i) = I;
    }
    // Maximization step
    for k=1, k<K, k++ {
        Xsubset = X(find(Xk==k));
        newMUs(k) = mean(Xsubset);
    }
    e = sum(newMUs-MUs)/sum(MUs);
    n++;
    MUs = newMUs;
}

Fig. 7. Pseudocode for K-means algorithm

Mixture Model (GMM), a combination of a fixed number of K Gaussian distributions to probabilistically represent the data. This process is significantly more computationally intensive than K-means, but results in a spatial parameterization of clusters in the form of the covariance matrix, \( C \).

In the generalized EM algorithm the two steps in the convergence process are:

- **Expectation**, in which point labels are assigned to a fixed model (typically a Mixture of Gaussians, though any proper distribution can be used)

- **Maximization**, in which the model is optimized to fit the current fixed point assignments such that the log-likelihood of the distribution model is maximized.

The result of the maximization step for a Gaussian distribution yields the covariance matrix, \( C \). For two-dimensional point data, \( C \) is a \( 2 \times 2 \) diagonal matrix, as the calculation of the covariance matrix is analogous to an Inner Product Matrix of 2-D data. This matrix has the following properties:

- Bilinear, such that \( C(ax + by, z) = aC(x, z) + bC(y, z) \)
- Symmetric, such that \( C(2, 1) = C(1, 2) \)
- Positive semi-definite, such that \( C(1, 1) \geq 0 \) and \( C(2, 2) \geq 0 \)

Using the previous example data, the shape data (experimental covariance) of 25 example clusters yields insight into the shape, scope, density, and overlap of these crime hotspots, as in Figure 10.

3) **Correlation and Contextual Analysis:**

C. **Data Presentation**

In order to better visualize the different crime hotspots and trends, we designed a web-based graphical user interface. The user interface displays crime hotspots based on different crime types, hotspots based on different counties, cluster movement and allocation of certain crime types, trend comparison of different crime types, a special case study on London Riots and the correlation of different crime types. These sections are described in detail below in order of their appearance on the website.

1) **Crime Hotspots for Various Crime Types:** This section displays the different crime hotspots for all the crime types for whole of Merseyside. A user can select a particular crime type and see several different analysis for that particular crime type. An example of this visualization tool is shown in Figure 11.

This section allows a user to understand how a particular crime type has evolved for whole of Merseyside. The analysis includes the following:

- **Hotspot Visualization on Google Maps:** The top 7 hotspots for that particular crime type in whole of Merseyside are displayed on Google Maps. These
hotspots are color coded where the darkest color shows the borough with the maximum occurrence for that particular crime type. A user can further click on individual borough to obtain further information about the total occurrences of that particular crime type for that borough. The boroughs are labeled using the English alphabets from A till F and are further classified into $A_1, A_2, \ldots, A_n$.

- **Trends in Hotspots**: A line graph on the right-hand side of the UI shows the incident occurrence of that particular crime type over time for each of the 7 hotspots. Using this, a user can easily visualize how a particular crime type has evolved over time.

- **Correlation with other crime types** This graph shows the correlation of the selected crime type with the other crime types. This correlation has been obtained over time and allows a user to visualize if one crime type can have a positive/negative effect on some other crime type.

- **Live Tweet Feeds**: This section shows the live Tweets for that particular crime type from the whole of Merseyside. This allows a user to understand what people are talking about for a particular crime type. Interesting examples in this case are the tweets on Mugging, anti-social behavior, weapons, drugs and burglary, where one can see quite opinionated tweets.

- **Live News Feeds**: This section shows the live news feeds for the selected crime type for whole of Merseyside. This allows a user to understand the current facts that are being reported for that particular crime type.

2) **Crime Hotspots based on different Areas**: This section displays the different hotspots and trend analysis for all the crime types combined for a particular Metropolitan Borough. This allows a user to garner knowledge about a particular metropolitan and understand which areas are safe and which are unsafe and how active is the police and judiciary for those areas. An example of this visualization tool is shown in Figure 12.

The analysis includes the following:

- **Hotspot Visualization on Google Maps**: The top 5 hotspots for the selected metropolitan is displayed on Google maps. These hotspots are generated by combining all the crime types together. As before even these hotspots are color coded, where the dark color means the hotspot with maximum number of occurrences. A user can select these hotspots to identify the crime count for each of them.

- **Trends in the Hotspots**: A line graph shows how these
top 5 hotspots have evolved over time and is there a trend associated with these hotspots.

- **Prosecution Data Analysis:** A pie chart shows the prosecution data for all the crime types for the selected metropolitan area. The prosecution data contains information about the current state of prosecution and whether the reported crime was solved or not. Surprisingly for most of the metropolitans a big percentage of the reported crime goes unsolved thereby making it alarming.

- **Live Tweet Feeds:** This section shows the live tweet feeds specific to that particular metropolitan area and mentioning crime. This allows a user to understand the sentiment of the people living in that metropolitan area about the crime.

- **Live News Feeds** This section shows the latest news on crime from the selected metropolitan area and helps a user to understand the current happenings on crime for that region.

3) **Crime Trends over Time:** This section compares the different crime types trends for the year 2012 with the crime type trends from the year 2011. This gives us very vital information about how the different crime types have evolved as compared to their numbers from last year. An example of this visualization tool is shown in Figure 13.

A few interesting observations from these trends are as follows:

- For almost all the crime types except shoplifting, the crime has decreased as compared to last year data. The blue line shows the data from 2011 and the orange line shows the data from 2012.

- For most of the crime type, we can see seasonal effects where the crime type decreases in September and then increase again in October.

- For most of the crime types, there is a significant difference between the crimes reported in August 2011 as compared to August 2012, because of the London Riots.

4) **Cluster Analysis of Certain Crime Types:** We did clustering for certain crime types using K-means to identify the movement of clusters and hotspots over time. This section displays the resulting clusters for these crime types and their movement over time. The crime types used for these clustering includes burglary, anti-social behavior and total crime. A user can view the clusters over time and then can select a particular cluster to see its behavior and movement temporally, shown in Figure 14.

Within this visualization module, various sections are
available for the user to view and draw inference from:

- **All Clusters assignment over time** This section displays the clusters assigned spatially over time for each of the month displayed on Google Maps. Clicking on each of the clusters generate vital information about the cluster which includes cluster number, cluster area and the number of incidents reported for a particular crime type.

- **Tracking particular cluster temporally** A user can click on a particular cluster to view further information about it, which includes the number of point assignments to that cluster and the correlation matrix. This information allows a user to visualize how stable a cluster hotspot is over time and the trend (increasing/decreasing/stable) shown by that particular cluster.

V. Results

As a case study, all crime instances categorized as burglary were analyzed over a period of 21 months, spanning from December 2010 through August 2012. This data was incorporated into a three-part analysis:

1) Clustering was performed on the global dataset to establish overall centroid locations
2) Secondary clustering was performed month-to-month to establish cluster shape parameters
3) Cluster parameters are tracked and analyzed over time to establish spatiotemporal crime trends

A. Global Clustering, Determining K

To establish global centroid locations, all crime from the dataset was combined into a single spatial array of instances. These instances were clustered with a range of $K$-values from 2 to 55, and the Sum of Squares Error (SSE) was calculated for each clustering result.

$$SSE[y] = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

The SSE of a clustered dataset is an approximate indicator of the global point assignment error – each point is measured against its assigned centroid, and this squared Euclidian distance is summed over the entire dataset.

Of note, because the convergence result for K-means is not necessarily identical between iterations of the algorithm for various values of $K$, the SSE result is not always monotonically decreasing. That is, for some values of $K$, the SSE for a particular convergence result may be higher than the SSE for $K-1$. To address this, the threshold for selecting $K$ is not based on a second-derivative of the error function, but instead selected when the SSE stabilizes to within some user-specified stability range of a final value.

The relationship between SSE and $K$ allows the identification of a value of $K$ at which no significant reduction in SSE can occur. Figure 15 shows the plotting of SSE against various values of $K$. In this case, the SSE stabilizes to within 95% of a final value above $K = 35$. The results indicate that there exist negligible gains in SSE for $K$ values above 35. So, $K = 35$ clusters was chosen for a segmentation value for this dataset.

B. Expectation Maximization and Shape Parameterization

The Expectation Maximization (EM) algorithm was applied to each month of data from the burglary dataset. This algorithm fits the data to a Gaussian Mixture Model (GMM) with a specified number of clusters. The EM algorithm for GMMs works in several key steps:

The mixture model assumes an initial mean, covariance matrix, and total probability for each cluster, and estimates the partial probability of each data point relative to the cluster mean:

$$P(x|\mu_k, S_k) = S_k^{-1/2} \left( \frac{1}{2\pi} \right)^{d/2} e^{-\frac{1}{2}(x-\mu_k)^T S_k^{-1}(x-\mu_k)}$$

Given base probabilities for each cluster, the Total Law of Probability yields:

$$P(x|\mu_k, S_k, k) = P(x|\mu_k, S_k)P(k)$$

Finally, using Bayes law, the mean and covariance parameters can be obtained and the total log-likelihood of the clusters calculated. The Expectation Maximization algorithm iterates until the sum of log-likelihoods between iterations stabilizes to within a user-specified tolerance. The shape parameters (covariance matrices and means) obtained from each cluster are stored for each month, and tracked over time. In addition, the number of points assigned to each cluster are also tracked, yielding a running estimate of $P(k|x; t)$. 
C. Cluster Analysis: Case Studies

In order to detect trends and patterns within the data, several parameters were monitored across all months of the dataset.

To analyze the spatial distribution of each cluster with respect to time, the covariance elements were plotted as 3-D points. This is made possible by the diagonal symmetry of the covariance matrices, such that $C(2; 1) = C(1; 2)$. By plotting the covariance matrices in this way, it is possible to identify months of data for each cluster which quantitatively differ from normal.

Additionally, the number of crime incidents assigned to each cluster for each month was also tracked over time. This point assignment value indicates the approximate “level” of crime (of a specific type) in a geographic region for a fixed time period. Monitoring this value over time allows the analysis of factors such as local stability in crime rates, upward or downward crime trends, and crime densities (normalized to unit area).

This area-normalized value is found to correlate well with regional boundaries, as well as borough and postal code locations, explained further in the case studies in Section V-D.

The paper describes a mechanism for identifying correlation between crime patterns and social media trends. It uses several clustering mechanisms and data correlation matrices to identify a statistical relationship between the two. It aims at finding the spatial and temporal correlations in particular. It works on crime data collected from uniform crime reports and the social media data collected from blogs, news feeds, facebook, twitter, etc. The collected data is cleaned first to remove the noise and then analyzed to identify trends and relationships. These trends and relationships are displayed on a web-based system using google maps and several web-based chart libraries in an interactive manner to show the trends and predict events in future.

1) Case Study: Spatially Anomalous Behavior: For example, cluster 35 (centered at E 343630.8979591837, N 384038.0943877551) contains an anomalous point which differs from the principal spatial distribution of the data. Figure 16 shows this plot, with anomaly highlighted.

This anomaly occurs during month 4 of the dataset, in March of 2011, in Garston (a southeastern suburb of Liverpool, UK), shown in Figure 17. This type of spatial anomaly is indicative of a particular month in which the shape of the cluster was determined to be significantly different from normal”. This may indicate a sudden shift in crime, or may simply indicate that the point assignments for that month did not produce a hotspot with similar spatial characteristics. Contextual analysis and the incorporation of social media data is necessary to further analyze this type of anomaly. For example, the crime throughout Merseyside as a whole increases during the month of March 2011 (shown in Figure 18 in Section V-C2). While the cause for this increase is unclear, the increase itself may serve as a contextual reason for the anomaly in the spatial parameters for crime in Garston for that month.

2) Case Study: Incident Assignment Trend: Additionally, the $N$ values for each cluster (number of points assigned) were also tracked across time. Figure 18 shows a comparison of these tracking results for several clusters (5; 10; 15; : : : ; 35) for all 21 months of the dataset.

Using these clustering results, it is possible to predict trends in crime for a specific geographic region. This allows for better allocation of law enforcement resources and other community programs aimed at reducing crime. For example, cluster 15 (E 327064.1794871795, N 382806.7333333333) centers around Pensby, a small town southwest of Birkenhead and Bebington. Crime in this region is relatively stable, both spatially and temporally. As Spatial clustering for this region shows a tightly-grouped covariance, with little variation from
Burglary in this region appears to be trending downward, with stable spatial characteristics indicating that crime is not spreading into surrounding areas. It is reasonable to assume that the allocation of law enforcement resources in this area is appropriate to the regional level of burglary.

3) Case Study: Incident Assignment Instability: By contrast, cluster 26 (E 351666.4141078838, N 395951.3502074689) demonstrates a high degree of variability in crime density, with little rotational variability in cluster shape. Centered in the city of St. Helens, due east of Liverpool, this cluster represents a major metropolitan area within the county of Merseyside. Figure 20 shows the spatial covariance tracking of this cluster; little variability can be seen in the off-diagonal elements of the covariance matrix. However, the diagonal elements span a distance of nearly an order of magnitude larger than most comparable clusters. Combined with the high degree of variability in cluster assignments (Figure 21), the results indicate an unstable region with unpredictable burglaries. It may be beneficial for law enforcement to study this region in the context of social media in order to better understand the causes (and best methods to address) burglary in St. Helens.

D. Other Analyses: Case Studies

1) Case Study: London Riots: The data we have collected is from December 2010 till October 2012. Luckily for us this period also includes the famous London Riots which occurred in August 2011. The riots started because of the police shooting of a 29-year old Mark Duggan which led to a country wide protest and involved several instances of looting and arson. A visual breakdown of the analysis results is shown in Figure 22.

We studied these particular incidents in greater detail from several different aspects. These aspects are divided into five major areas as follows:
Fig. 22. A case study breakdown of the London Riots from August, 2011

- **Arrests in Merseyside based on Age** We classified the arrests made for the different crime types based on the age of the person arrested. Almost 40% of the arrests made involved juveniles and a lot of them were arrested for public disorder.

- **Arrests in Merseyside based on Gender** We classified the arrests made for the different crime types based on the gender of the person arrested. More than 95% of the arrests involved males.

- **Arrests in Merseyside based on Ethnicity** The arrests were classified based on the ethnicity. The major percentage of the arrested were whites followed by blacks.

- **Crime throughout the week** The crime is further segregated into the different days of that week. This graph gives user a very clear idea that the peak of the protest was on 8th and 9th August because that's when the maximum arrests were made.

- **Crime by Offense** Finally the arrests were classified into different crime types to understand what was the most prevalent crime incident being reported during that week. Damage to Public property contributes to the major chuck of the arrests made followed by violence against person.

2) **Case Study: Crime Type Correlations:** Additional analysis was performed to correlate the different crime types both spatially and temporally. This allows visualization of how one crime type is affected/dependent on another crime type. The User Interface visualization tool for this case study is shown in Figure 23.

A few interesting correlations were found by correlating crime types temporally. These include:

- A strong correlation of drugs with weapons. The correlation coefficient of drugs and weapons is 0.75
- A strong correlation of burglary with robbery, vehicle crime and arson. The correlation coefficient of burglary and robbery is 0.88, the correlation coefficient of burglary and vehicle crime is 0.68 and the correlation coefficient of burglary with arson is 0.65
- At the same time burglary is negatively correlated with shoplifting, anti-social behavior and violent crime.
- Shoplifting and anti-social behavior shows a strong positive correlation with a correlation coefficient of 0.7.
- There is a strong correlation between arson and vehicle crime with the correlation coefficient of 0.68
- The biggest negative correlation is between robbery and anti-social behavior with a correlation coefficient of $-0.65$, followed by a negative correlation of robbery with shoplifting with a correlation coefficient of 0.58

3) **Case Study: Cluster Hotspot Rankings:** This section displays the rankings of the different clusters for certain crime type over time. Using this information a user can identify the number of point assigned to a particular cluster and the point assignments normalized by area. A few interesting observations are as follows:

- For all crime types, cluster 14, shows a significant amount of points assigned normalized by area
- For anti-social behavior cluster 14 and 18 shows a significant amount of points assigned normalized by area.
- For burglary, cluster 5 and 18 shows significant amount of points assigned normalized by area.

An example of this visualization tool from the User Interface is shown in Figure 24.
VI. CONCLUSION

The paper describes a mechanism for identifying and parameterizing crime hotspots, and establishing correlations between crime patterns. This study implements several clustering mechanisms and analysis tools to identify statistical relationships within the data, track crime hotspots parameters over time, predict crime trends in geographic locations, and both spatial- and temporal- correlative measures between various crime types.

The region of interest, Merseyside UK, has been the subject of prior studies. This previous information allows us to validate our results against existing approaches. Several key points of interest drawn from our analysis of the data reveal interesting trends (or single months) in which the data behave in a way that is quantitatively different from “normal”.

Finally, these results are formed into a User Interface which displays the collected information to a user through a web-based tool, implementing features such as crime demographic breakdowns, cross-sections of crime types in a geographic region, correlative measures between crime types in a geographic region, and cluster shape and assignment parameters over time. This information is displayed in a way that is both comprehensive and intuitive, allowing the user to easily navigate the large amount of data present in this study.

We believe that the above results strongly support our approach as an analysis tool for similar study in other street-level crime datasets, and will enable law enforcement agencies to draw more rapid and accurate conclusions about their own spatiotemporal crime parameters. This contextual information will result in more effective resource allocation, better community awareness, and hopefully a decrease in crime in general.

REFERENCES

## Appendix A
### Progress Update

<table>
<thead>
<tr>
<th>Task</th>
<th>Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P. Saraf</strong></td>
<td></td>
</tr>
<tr>
<td>Conduct literature review</td>
<td>In Progress</td>
</tr>
<tr>
<td>Research potential available social media data</td>
<td>Ongoing</td>
</tr>
<tr>
<td>Obtain social, demographic, and educational data for the UK</td>
<td>In Progress</td>
</tr>
<tr>
<td>Correlate crime data with additional regional information</td>
<td>Not Complete</td>
</tr>
<tr>
<td>Assist in the development of UI</td>
<td>In Progress</td>
</tr>
<tr>
<td><strong>M. Milo</strong></td>
<td></td>
</tr>
<tr>
<td>Conduct literature review</td>
<td>Complete</td>
</tr>
<tr>
<td>Research potential methodology</td>
<td>Complete</td>
</tr>
<tr>
<td>Obtain street-level crime dataset</td>
<td>Complete</td>
</tr>
<tr>
<td>Develop and implement k-means algorithm</td>
<td>Complete</td>
</tr>
<tr>
<td>Refine and finalize k-means algorithm</td>
<td>Complete</td>
</tr>
<tr>
<td>Develop and implement EM/GMM algorithm</td>
<td>Complete</td>
</tr>
<tr>
<td>Preliminary analysis of shape parameters (covariance matrices)</td>
<td>Complete</td>
</tr>
<tr>
<td>Data extraction, parsing, monthly segmentation by crime type</td>
<td>Complete</td>
</tr>
<tr>
<td>Clustering and EM application to Burglary dataset</td>
<td>Complete</td>
</tr>
<tr>
<td>Trend analysis and anomaly detection for Burglary set</td>
<td>Complete</td>
</tr>
<tr>
<td>Contextual analysis and geospatial indexing by cluster / city</td>
<td>Complete</td>
</tr>
<tr>
<td>Checkpoint 2 analysis, figure generation</td>
<td>Complete</td>
</tr>
<tr>
<td>Furnish results to other team members for social media analysis</td>
<td>Complete</td>
</tr>
<tr>
<td><strong>L. Moonga</strong></td>
<td></td>
</tr>
<tr>
<td>Conduct literature review</td>
<td>Complete</td>
</tr>
<tr>
<td>Obtain regional geographic information (boundaries)</td>
<td>In Progress</td>
</tr>
<tr>
<td>Conduct exploratory GIS Analysis</td>
<td>Ongoing</td>
</tr>
<tr>
<td>Research potential sources of data to correlate with crime data</td>
<td>In Progress</td>
</tr>
<tr>
<td><strong>S. Richards</strong></td>
<td></td>
</tr>
<tr>
<td>Conduct literature review</td>
<td>Complete</td>
</tr>
<tr>
<td>Obtain regional geographic information (aggregated demo-</td>
<td>Complete</td>
</tr>
<tr>
<td>graphic/landuse data)</td>
<td></td>
</tr>
<tr>
<td>Conduct exploratory GIS Analysis</td>
<td>Complete</td>
</tr>
<tr>
<td>Research potential statistical techniques to analyze data</td>
<td>Complete</td>
</tr>
<tr>
<td>Develop visual exploration of time series data</td>
<td>Complete</td>
</tr>
<tr>
<td>Generate Typeset documents</td>
<td>Complete</td>
</tr>
<tr>
<td><strong>T. Bhattacharjee</strong></td>
<td></td>
</tr>
<tr>
<td>Conduct literature review</td>
<td>In Progress</td>
</tr>
<tr>
<td>Research existing web applications</td>
<td>In Progress</td>
</tr>
<tr>
<td>Design, develop, and implement web-based UI</td>
<td>In Progress</td>
</tr>
</tbody>
</table>