Quantifying Behavioral Data Sets of Criminal Activity

Jameson Toole
Department of Physics,
University of Michigan,
Ann Arbor, Michigan 48109, USA

Nathan Eagle
The Santa Fe Institute,
1399 Hyde Park Rd, Santa Fe,
New Mexico 87501, USA

Joshua Plotkin
Department of Biology and Program in Applied Mathematics and Computational Science,
The University of Pennsylvania,
Philadelphia, PA 19104, USA

Abstract

With the increased availability of rich behavioral data sets, we present a novel combination of tools to analyze to analyze this information. Using criminal offense records as an example, we employ cross-correlation measures, eigenvalue spectrum analysis, and results from random matrix theory to identify spatiotemporal patterns. Finally, with multivariate autoregressive models, we demonstrate a possible source of structure within the data. instructions.

Introduction

There is an enormous amount of behavioral data generated and stored by billions of individuals across countries and cultures. It has become necessary to develop novel quantitative tools to analyze this immense and rich stream of information. The goal is to use the data in order to gain a better understanding of the systems that generate it. This will inform fields from economics to sociology as well as provide policy makers with critical answers that may be used to better allocate scarce resources or implement beneficial social programs. In this paper, we present a novel combination of tools and analytical techniques that may be used to identify patterns and signals that capture fundamental dynamics of a social system. Cross and auto-correlation measures are combined with autoregressive models and results from random matrix theory to examine lead-lag relationships in behavioral time series.

The data set used in the course of this study is made up of criminal activity within the City of Philadelphia during the year 1999. For the roughly two hundred thousand crimes reported, we examine spatial, temporal, and incident information. The goal of our analysis is to explore the spatiotemporal dynamics of criminal behavior with the hope of identifying patterns that may be useful in predicting and preventing future criminal activity.

Existing work from the fields of criminology, sociology, psychology, and economics tends to explore relationships between criminal activity and socioeconomic variables such as education, community disorder, ethnicity, ect. (Weisburd, Bruinsma, and Bernasco 2009), (Lafree 1999). In general, constraints on the availability of data meant these studies were limited to aggregate statistics for large populations and vast geographic regions. Wilson and Kelling's article in the March, 1982 edition of The Atlantic popularized "Broken windows" and "social disorganization" theories, for the first time, explicitly introducing flow and dynamics into crime research. These theories proposed that crime was a consequence of urban decay and lack of community ownership in neighborhoods (Kelling and Wilson 1982). Neglected areas not only attract criminals (the neglect is a sign of lack of police presence), but also act as a feed-forward mechanism by damaging community moral.

More recently, attempts have been made to study crime on the neighborhood level, exploring crime "hot spots" (Sampson, Raudenbush, and Earls 1997). Some studies have even shifted focus from high crime areas to high risk people, tracking individuals for a period of time and assessing their propensity to commit crime and its relationship to various socioeconomic indicators (Krivo and Peterson 2000). These studies tend to be small in size and very labor intensive, requiring that neighborhoods be surveyed and tracked for long periods of time.

In this research we address a gap. Statistical methods have been used to characterize large, aggregate data sets over long periods of time, while sociological studies have been performed at micro scales. There remains a need for a high resolution quantitative analysis of large crime data sets. Using offense reports generated by a police department, we explore how crime here and now affects crime there and then, while also focusing on building a general set of tools to analyze behavioral data sets for spatiotemporal systems.

Data

Representing nearly all reported crimes within the City of Philadelphia, roughly 200,000 criminal events were recorded at nearly 37,000 unique locations. For each event, information is available about the time, place, and type of offense.

The spatial resolution of this data is high enough that a block/neighborhood analysis of crime is possible. Simply plotting the geocoded events reveals features of the city such as the street-grid, parks, bridges, rivers, ect. (FIG. 1). While the time of each report is known to within the hour, offenses are generally aggregated to daily or weekly counts, ensur-

ing that time series are sufficiently populated with events. A time series displaying citywide theft-related crimes reveals weak seasonal trends as well as singular events such as holidays (FIG. 2). Finally, offense report statistics organized by type of crime can be explored (FIG. 3).

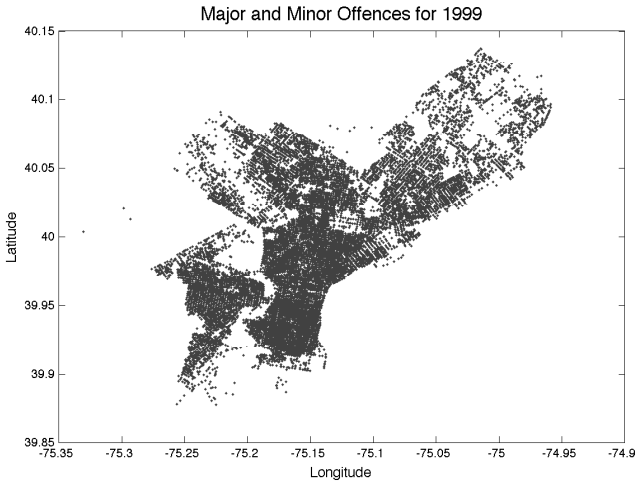

Figure 1: All crimes, major and minor, throughout the city of Philadelphia during the year 1999. Geographic features of the city, such as rivers, parks, bridges, ect., are immediately visible.

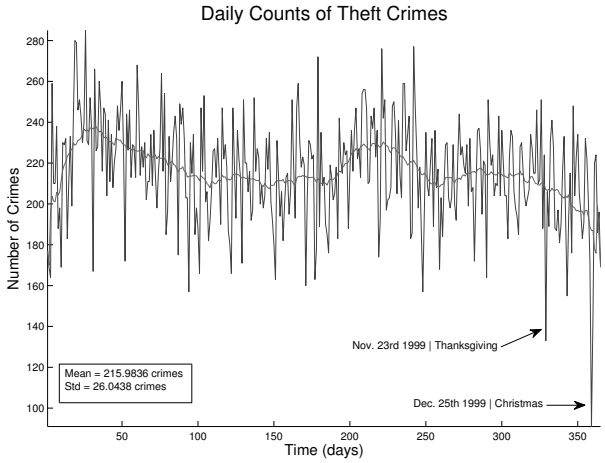

Figure 2: A time series plot of theft crimes. Significant outliers can be identified as holidays such as Thanksgiving and Christmas. Smoothing the data (the red/central line) reveals weak seasonal trends.

Methods

Conditioning the Data

Our goal is to quantitatively study a behavioral data set. Behavioral information must be transformed into variables that can be manipulated numerically. While time and place readily lend themselves to such analysis, the type of crime being reported must be inferred from its police description. Crimes were aggregated into six broader categories by parsing reports for keywords as described in Table 1. Aggregation ensures that there are sufficient numbers of events to populate time series, while still making use of nearly 75% of the available data. A lattice is laid over a map of the city and crimes are aggregated to the nearest grid point. In general, any mesh size can be used; our analysis is performed with 80-120 grid locations across the city. This spacing roughly corresponds to neighborhood and census track sizes, allowing for possible comparison to demographic census tract data. For every lattice point, a time series is constructed for each category listed above. In most cases, the entries of time series are the number of crimes at that location during a 24 hour period, though weekly time series are constructed in certain instances. Finally, time series are normalized to have zero mean and unit variance. Overall, the time series produced are stable and stationary, having roughly constant mean and variance over the year. Any seasonal effects are weak at best.

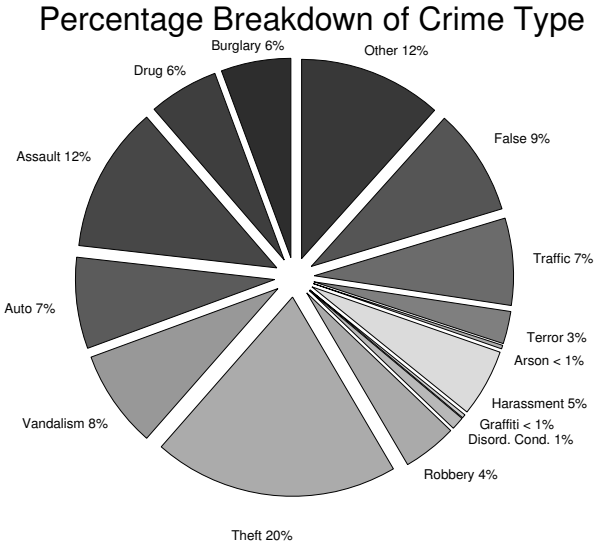

Figure 3: A percentage breakdown of different crimes based on incident reports.

Table 1: Categorical groupings of different crime types.

Category	Offenses Included	Crimes (%)
All	all reported offenses	211606 (100%)
Automobile	auto theft, major traffic	16005 (7.6%)
Theft	burglary, robbery, auto	80290 (38%)
Violent	assault, homicide, gun	29908 (14%)
Vandalism	vandalism, graffiti	17880 (8.5%)
Drug	possession, sale, DUI	11642 (5.5%)

Tools and Techniques

Using the conditioned data, we will develop analytical tools to achieve the following:

1. Identify both spatial and temporal relationships.

2. Assess the significance of these relationships.
3. Provide insight into their source.

To address these goals, we combine time series analysis with results from random matrix theory to quantify the magnitude and significance of correlation in the data. In the process, we draw upon knowledge of similar problems found in neuroscience (correlating spike-train data) and financial economics (finding signal in noisy time series data) (Kamiński et al. 2001), (Mayya and Santhanam 2007), (Tumminello, Lillo, and Mantegna 2008), (Laloux et al. 1999) . Finally, we present a reasonable source explaining observed patterns and suggest further research that may reveal deeper trends.

Correlation In Time: We begin analysis by looking for correlations in time. We first select two time series, $\mathbf{y}_1$ and $\mathbf{y}_2$, from the conditioned data. These time series may come from two different lattice points (so as to correlate points in space), or correspond two different crime types at the same lattice point (assessing lead-lag relationships between specific kinds of criminal activity).

The cross-correlation, $r_{1,2}$, is a measure of similarity between a pair of time series. Mathematically, this quantity is defined as the expectation of the inner product between the two time series: $r_{1,2} = E[\langle \mathbf{y}_1, \mathbf{y}_2 \rangle] = \sum_{t=1}^{n} y_1(t) y_2(t)$. Similarly, it is possible to determine lagged correlation by shifting one series by a number of lags, m. The lagged cross-correlation, $r_{1,2}(m)$, is given by modification to the previous formula, $r_{1,2}(m) = \sum_{t=1}^{n} y_1(t+m) y_2(t)$.

For our normalized time series, cross-correlation values lie between -1 and 1, where $r_{1,2} = 1$ is corresponds to exact correlation between two time series. A cross-correlation sequence is defined as the sequence of cross-correlation values over a range of lags. Examining the cross-correlation sequence for two time series, we can identify the existence of a significant relationship as well as quantify its power over a number of lags. Not only can these measures detect the flow of crime from one area to another, they can also quantify its speed and direction. An example of such measures can be found in FIG. 4.

Correlation in Space: Comparing time series for all pairs of locations across the city, we form a $K \times T$ matrix, $\mathbf{Y}$, where K is the number of time series we wish to correlate and T is the length if each time series. The delayed correlation matrix for a specific lag m, $\mathbf{C}(m)$, is then constructed by matrix multiplication $\mathbf{C}(m) = \frac{1}{T}\mathbf{Y}\mathbf{Y}^T(m)$, where ^{T} is a regular matrix transposition. The elements of $\mathbf{C}$ are given by $C_{ij}(m) = \sum_{t=1}^{T} y_i(t) y_j(t+m)$. Note that $m = 0$ corresponds to zero lag.

For example, we wish to test for correlations in drug related offenses between different neighborhoods across time. Conditioning the data as described above, time series are constructed for 35 lattice points (neighborhoods) and the cross-correlation matrix is constructed (FIG. 5). Entry C_{ij} of this matrix represents the cross-correlation between the time series of drug related crimes from locations i and j. Examining this matrix we first note that no patterns or regions of high correlation are immediately visible. The labeling of neighborhoods is such that locations i and $i+1$ are also close spatially. The unstructured correlation matrix suggests that neighborhood crime levels may not be correlated spatially. This example, though, does not consider any lagged correlations that may exist between locations. Constructing matrices for lagged cross-correlation of up to 30 days (1 month), however, reveal similar results. We do not find any immediate spatial correlation structure or flows in the data.

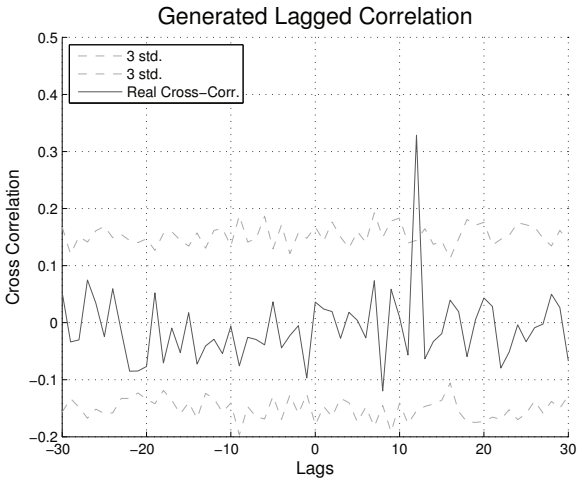

Figure 4: The cross-correlation sequence of two generated time series. Time series $y_1(t)$ was generated randomly while y_2 is a linear combination of values of $y_1(t-12)$ and white noise. The blue (solid) line represents the actual cross-correlation sequence while the green (dotted) lines represent 3σ significance tests.

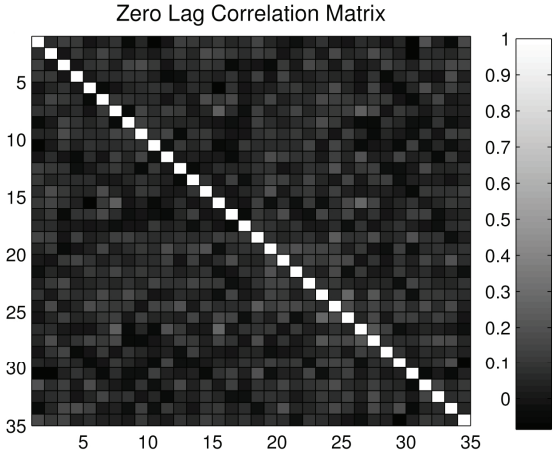

Figure 5: The zero lag correlation matrix for drug related crimes. There appears to be little spatial correlation and a lack of high correlated locations.

Correlation by Type: Cross-correlation methods may also be used to look for relationships between crimes. We may ask if an increase in theft related crimes leads to violent crimes in the future. For a given node, we create a time

series for each type of crime. Next, we construct the cross-correlation sequence for this pair of series across a number of lags (in most cases lags up to 30 days were included). To visualize these correlations we create a matrix where each column represents the cross-correlation sequence for a given location (FIG. 6).

As an example, we have included automobile thefts in both the "Thefts" category and the "Automobile" category. Unsurprisingly, we see significant correlation between the two crime types at exactly zero lag. The lack of significant correlation for other time lags indicates no other significant relationships where theft in one location leads to violence in that same location at a later time.

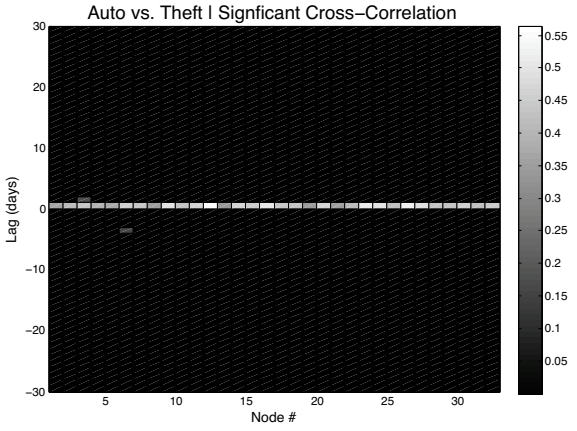

Figure 6: A matrix displaying significant lagged cross correlations between automobile crimes and theft crimes. Because automobile crimes are counted in both categories, we find correlation at zero lag, but almost no other significant relationships.

Significance: Across each crime category, we find few visible spatial or temporal correlations. Given that no immediate structure is present, our goal is to assess the significance level of cross-correlation values to differentiate between real and random connections. To achieve this, we create a null model for each pair of time series by randomizing the series and computing a new cross-correlation value. Repeating this process 500 times, we construct a distribution of cross-correlation values from which confidence intervals can be constructed. If the cross-correlation between the original time series deviates from the random distribution at a given confidence level, we consider it significant. Performing this analysis, however, we find no significant correlation structure in the criminal data.

Given the random nature of our findings, we again turn to work done on analogous systems. Similar problems involving noisy time series are routinely found in fields such as financial economics (markets. stocks, equities, ect.) and climate forecasting (Tumminello, Lillo, and Mantegna 2008), (Laloux et al. 1999). Work in these areas suggests that the correlation structure of systems can be characterized by examining the eigenvalue spectra of correlation matrices. Much success has been found testing the significance of these metrics using results from random matrix theory (RMT).

To test for non-random structure in our correlation matrices, we consider two related groups of matrices, Gaussian and Wishart. Entries of a Gaussian matrix are drawn from a standard normal and a Wishart matrix, $\mathbf{W}$, is formed by matrix multiplication of a Gaussian matrix, $\mathbf{G}$, and its transpose, $\mathbf{W} = \mathbf{G}\mathbf{G}^T$ (Edelman 1988). The key observation is the direct analogy between formulation of the Wishart matrix and cross-correlation matrix. We use these random matrices as null models to our cross-correlation measures.

Various analytical results for the distribution of eigenvalues of a random Wishart matrix can be found in (Edelman 1988), (Sengupta and Mitra 1999), (Utsugi, Ino, and Oshikawa 2003). The eigenvalue density, $\rho(\lambda)$, is defined as the number of eigenvalues below λ. Given a correlation matrix whose entries are drawn from the standard normal distribution, the eigenvalue density as K and T go to infinity is given by the Marcenko-Pastur Law,

$$\rho(\lambda) = \frac{Q}{2\pi} \frac{\sqrt{(\lambda_{max} - \lambda)(\lambda - \lambda_{min})}}{\lambda} \qquad (1)$$

where $Q = T/K \geq 1$ and $\lambda_{min}^{max} = 1 + 1/Q \pm 2\sqrt{1/Q}$ (Laloux et al. 1999). In other words, we can establish significant correlation structure by comparing eigenvalue spectra from the data to those of a random null model. If we find eigenvalues significantly outside theoretical thresholds, we can conclude there is signal buried in the data. For example, the largest eigenvalue (and corresponding eigenvector) in the case of financial data is identified as the "market" factor, having equally weighted components. Mayya *et. al.* have obtained similar analytical results for lagged cross correlation matrices (Mayya and Santhanam 2007).

Identifying Patterns: Examining the eigenvalue spectrum of our data, we do find a weak signal. The spectra corresponding to the correlation matrix of drug related crimes displayed above is shown in FIG. 7. While the majority of eigenvalues cannot be distinguished from noise, there does exist a large significant eigenvalue.

Next we examine the spectra for a series of lagged correlation matrices. Plotting the magnitude of the largest eigenvalue for each lagged correlation matrix and comparing this to the largest value expected from random data, we see a strong cyclic signal with a period of 7 days. (FIG 8). This analysis suggests that significant correlation structure is present on a weekly cycle.

Recreating Patterns We now address the source of this correlation. Again borrowing from results in financial economics, we adopt use of the inverse participation ratio (IPR) to examine the component structure of the significant eigenvectors. The IPR of a vector is given by $IPR(\vec{\nu}_i) = \sum_{j=1}^{K} |\nu_{ij}|^4$ (Biely and Thurner 2006). A large IPR implies that only a few components contribute to the eigenvector, while a small IPR indicates participation of many components. It is possible to determine clustering structure from such analysis. For example, in financial data, the eigenvector corresponding to the large "market" eigenvalue has

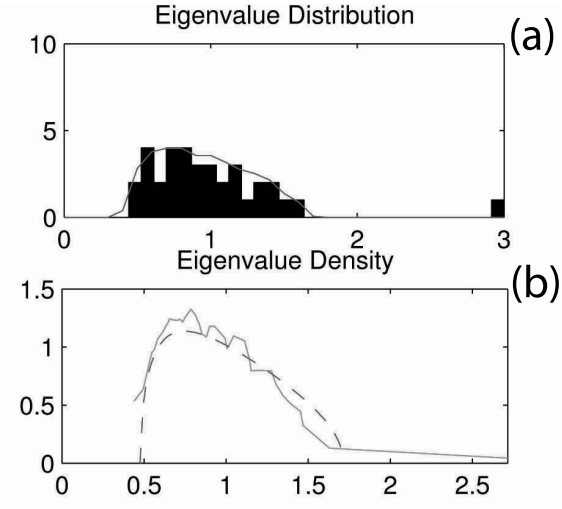

Figure 7: (a) Only one eigenvalue, $\lambda_1 = 3$, can be differentiated from the noise indicated by the red (solid) line. (b) The green (solid) curve is the eigenvalue density of the actual matrix spectra while the red (dashed) curve is the theoretical prediction from eqn. (1).

a low IPR, identifying itself as a force that affects all stocks equally. Other eigenvectors, however, with larger IPRs have components that correspond to various sectors of the market (Biely and Thurner 2006). For crime data, these components correspond to locations across the city so a cluster of eigenvector components would correspond to a cluster of neighborhoods.

Examining the IPRs for significant eigenvectors in lagged correlation matrices, our results show that the eigenvector corresponding to the largest eigenvalue has a low IPR and can thus be interpreted as a "market" force. For the remaining significant eigenvectors, we find that they too have low IPRs, suggesting there is little clustering or community structure (FIG. 9).

Ruling out significant spatial clustering, we suggest possible temporal sources for the patterns observed in the data. While multivariate autoregressive models quickly become intractable if attempting to include crime levels at many locations across the city, they do provide some insight. Regressing citywide drug offenses on day of the week reveals significant correlation. Considering only what day it is, we are able to account for nearly 60% of the variance in daily drug offenses. Inference from the coefficients in Table 2 is subtle. With Sunday being the omitted group, coefficients on dummy variables corresponding to the day being Monday-Saturday are interpreted as the change in criminal activity between Sunday and that particular day of the week. Thus we conclude Sunday's have the lowest drug-related crime rates while the middle of the week (Tuesday, Wednesday, and Thursday) show the highest. This is in sharp contrast to violent crimes, which show an increase on weekends, dropping during weekdays.

While differences in when types of crime occur are intriguing, we focus primarily on the weekly, cyclical pat-

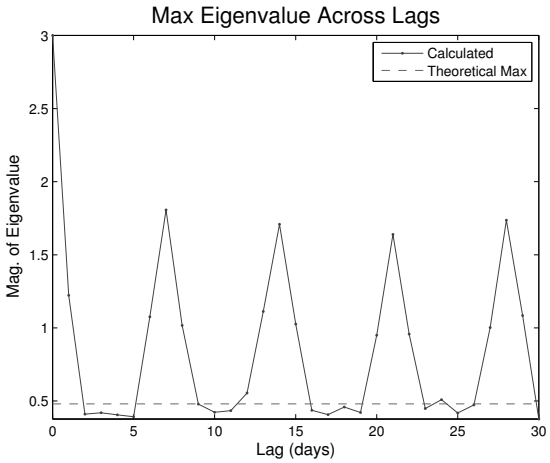

Figure 8: We plot the maximum eigenvalue of the delayed correlation matrix for each of 30 lags. For drug related crimes, we see a very clear periodicity at a frequency of 7 days (1 week).

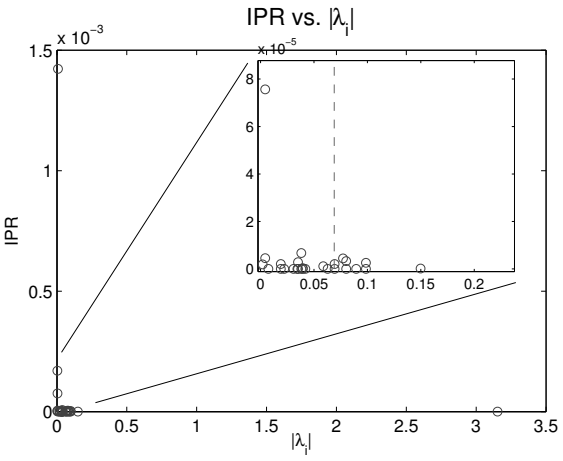

Figure 9: A plot of the IPR of the eigenvectors of the delayed correlation matrix for drug crimes with a 7 day lag.

tern it shows. Using the coefficients from the autoregressive models and some added noise, we construct a set of time series to mimic our original data. Calculating correlation matrices and their eigenvalue spectra, we are able to reconstruct the significant features of real crime data. Our generated model successfully reproduces the weekly spikes seen in the amplitude of the maximum eigenvalue of the symmetric lagged correlation matrices (FIG. 8), suggesting this citywide, weekly cycle is the major component driving correlation.

Summary and Conclusion

In this paper, we have presented a novel combination of tools that can be used to analyze behavioral data sets. Cross-correlation measures were used to construct correlation matrices, revealing spatiotemporal relationships involving human activity. Given a low signal-to-noise ratio, we adopted

Table 2: A regression of citywide drug and violent offenses on day of the week ($R^2_{drug} = .60$, $R^2_{viol} = .22$).

Day	Drugs Coeff. [95% Conf]	Violence Coeff. [95% Conf]
Sunday	-1.21 [-1.38, -1.04]	0.52 [0.29, 0.76]
Monday	.40 [0.16, 0.64]	-1.04 [-1.37, -0.71]
Tuesday	1.94 [1.73, 2.22]	-0.82 [-1.16. -0.49]
Wednesday	2.12 [1.88, 2.37]	-0.79 [-1.13, -0.46]
Thursday	1.87 [1.62, 2.12]	-0.94 [-1.28, -0.61]
Friday	1.39 [1.15, 1.63]	-0.37 [-0.71, -0.04]
Saturday	0.73 [0.48, 0.98]	0.24 [-0.09, 0.57]

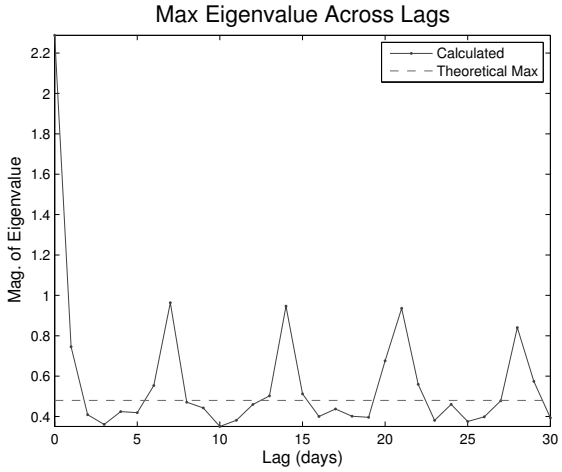

Figure 10: The maximum eigenvalue of the symmetric delayed cross correlation matrix for each of 30 lags. Generated data from regression coefficients captures the weekly periodicity of actual drug related crimes shown in FIG. 8.

results from random matrix theory to construct a suitable null model to construct significance tests. These tests revealed definite structure in the eigenvalue spectra of our correlation matrices. Finally, we present a method capable of generating observed patterns.

Given the large portion of crime rates that can be explained by regressing data onto the day of the week, it is possible that these results reflect police procedures such as scheduling more officers on Mondays than Sundays. Discrepancies in daily crime rates for different types of crime, however, may suggest different types of crime do represent very different behaviors. Another interesting result from our analysis is the lack of correlation between these different crime types. Broken windows and social disorganization theories postulate that an influx of minor offenses such as graffiti and vandalism might lead to an increase of more serious crimes such as assaults or gun violence. We find no evidence of this for spatiotemporal time scales probed here.

This is not to say, however, no relationship exists. We have only looked for interaction on time scales of up to 30 days. It may be that these types of flows happen on the monthly or yearly time scale. The length of our time series, however, limits us. Having recently acquired similar crime data for the entire decade, we hope to address this issue in future works.

We are confident, that these methodologies are capable of capturing patterns and dynamics within behavioral data sets and can help provide insight into the social systems that create them.

Acknowledgements

We would like to thank the Santa Fe Institute and the wonderful community of people who have helped use with their thoughts and ideas. We would also like to thank NSF Summer REU program for partially funding this research.

References

Biely, C., and Thurner, S. 2006. Random matrix ensembles of time-lagged correlation matrices: Derivation of eigenvalue spectra and analysis of financial time-series.

Edelman, A. 1988. Eigenvalues and condition numbers of random matrices. *SIAM J. Matrix Anal. Appl.* 9(4):543–560.

Kamiński, M.; Ding, M.; Truccolo, W. A.; and Bressler, S. L. 2001. Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernetics* 85(2):145–157.

Kelling, G. L., and Wilson, J. Q. 1982. Broken windows. *The Atlantic*.

Krivo, L. J., and Peterson, R. D. 2000. The structural context of homicide: Accounting for racial differences in process. *American Sociological Review* 65(4):547–559.

Lafree, G. 1999. Declining violent crime rates in the 1990s: Predicting crime booms and busts. *Annual Review of Sociology* 25(1):145–168.

Laloux, L.; Cizeau, P.; Bouchaud, J. P.; and Potters, M. 1999. Noise dressing of financial correlation matrices. *Physical Review Letters* 83(7):1467–1470.

Mayya, K. B. K., and Santhanam, M. S. 2007. Correlations, delays and financial time series. 69–75.

Sampson, R. J.; Raudenbush, S. W.; and Earls, F. 1997. Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* 277(5328):918–924.

Sengupta, A. M., and Mitra, P. P. 1999. Distributions of singular values for some random matrices. *Physical Review E* 60(3):3389–3392.

Tumminello, M.; Lillo, F.; and Mantegna, R. N. 2008. Correlation, hierarchies, and networks in financial markets.

Utsugi, A.; Ino, K.; and Oshikawa, M. 2003. Random matrix theory analysis of cross correlations in financial markets.

Weisburd, D.; Bruinsma, G. J.; and Bernasco, W. 2009. Units of analysis in geographic criminology: Historical development, critical issues and open questions. In Weisburd, D.; Bernasco, W.; and Bruinsma, G. J., eds., *Putting Crime in its Place: Units of Analysis in Geographic Criminology*. Springer.