Precise Event-level Prediction of Urban Crime Reveals Signature of Enforcement Bias

Victor Rotaru
University Of Chicago
Yi Huang
University Of Chicago
Timmy Li
University Of Chicago
James Evans
University of Chicago and Santa Fe Institute https://orcid.org/0000-0001-9838-0707
Ishanu Chattopadhyay (ishanu@uchicago.edu)
University of Chicago

Article

Keywords: enforcement bias, urban crime, crime rate

Posted Date: February 11th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-192156/v1

Version of Record: A version of this preprint was published at Nature Human Behaviour on June 30th, 2022. See the published version at https://doi.org/10.1038/s41562-022-01372-0.

Precise Event-level Prediction of Urban Crime Reveals Signature of Enforcement Bias

Victor Rotaru[1,3], Yi Huang[1], Timmy Li[1,3], James Evans[2,5,6] and Ishanu Chattopadhyay,[1,4,5]★

[1]Department of Medicine, University of Chicago, Chicago, IL 60637, USA
[2]Department of Sociology, University of Chicago, Chicago, IL 60637, USA
[3]Department of Computer Science, University of Chicago, Chicago, IL 60637, USA
[4]Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago, Chicago, IL 60637, USA
[5]Committee on Genetics, Genomics & Systems Biology, University of Chicago, Chicago, IL 60637, USA
[6]Santa Fe Institute, Santa Fe NM 87501, USA

★To whom correspondence should be addressed: e-mail: ishanu@uchicago.edu.

Policing efforts to thwart urban crime often rely on detailed reports of criminal infractions. However, crime rates do not document the distribution of crime in isolation, but rather its complex relationship with policing and society. Several results attempting to predict future crime now exist, with varying degrees of predictive efficacy. However, the very idea of predictive policing has stirred controversy, with the algorithms being largely black boxes producing little to no insight into the social system of crime, and its rules of organization. The issue of how enforcement interacts with, modulates, and reinforces crime has been rarely addressed in the context of precise event predictions. In this study, we demonstrate that while predictive tools have often been designed to enhance state power through surveillance, they also enable the tracing of systemic biases in urban enforcement—surveillance of the state. We introduce a novel stochastic inference algorithm as a new forecasting approach that learns spatio-temporal dependencies from individual event reports with demonstrated performance far surpassing past results (e.g., average AUC of $\approx 90\%$ in the City of Chicago for property and violent crimes predicted a week in advance within spatial tiles ≈ 1000 ft across). These precise predictions enable equally precise evaluation of inequities in law enforcement, discovering that response to increased crime rates is biased by the socio-economic status of neighborhoods, draining policy resources to wealthy areas with disproportionately negative impacts for the inner city, as demonstrated in Chicago and six other major U.S. metropolitan areas. While the emergence of powerful predictive tools raise concerns regarding the unprecedented power they place in the hands of over-zealous states in the name of civilian protection, our approach demonstrates how sophisticated algorithms enable us to audit enforcement biases, and hold states accountable in ways previously inconceivable.

THE emergence of large-scale data and ubiquitous data-driven modeling has sparked widespread government interest in the possibility of *predictive policing*[1–5]: predicting crime before it happens to enable anticipatory enforcement. Such efforts, however, do not document the distribution of crime in isolation, but rather its complex relationship with policing and society. In this study, we reconceptualize the process of crime prediction, build novel methods to improve it, and use it to diagnose both the distribution of reported crime and biases in its enforcement. The history of statistics has co-evolved with the history of criminal prediction, but also with the history of enforcement critique. Siméon Poisson published the Poisson distribution and his theory of probability in an analysis of the number of wrongful convictions in a given country[6]. Andrey Markov introduced Markov processes to show that dependencies between outcomes could still obey the central limit theorem to counter Pavel Nekrasov's argument that because Russian crime reports obeyed the law of large numbers, "decisions made by criminals to commit crimes must all be independent acts of free will"[7].

In this study, we conceptualize the prediction of criminal reports as that of modeling and predicting a system of spatio-temporal point processes unfolding in social context. We report a fundamentally new approach to predict urban crime at the level of individual events, with predictive accuracy far greater than has been achieved in past. Rather than simply increasing the power of states by predicting the when and where of anticipated crime, our new tools allow us to audit them for enforcement biases, and garner deep insight into the nature of the dynamical processes through which policing and crime co-evolve in urban spaces.

Classical investigations into the mechanics of crime[8–10] have recently given way to event-level crime predictions that have enticed police forces to deploy them preemptively and stage interventions targeted at lowering crime rates. These efforts have generated multi-variate models of time-invariant hotspots[11–13], and estimate both long and short term dynamic risks[1–3]. One of the earliest approaches to predictive policing is based on the use of epidemic-type aftershock sequences (ETAS)[4,5], originally developed to model seismic phenomena. While these approaches have suggested the possibility of predictive policing, many achieve only limited out-of-sample performance[4,5]. More recently, deep learning architectures have yielded better results[14]. Machine learning systems, however, are often black boxes producing little insight regarding the social system of crime and its rules of organization. Moreover, the issue of how enforcement interacts with, modulates and reinforces crime has been rarely addressed in the context of precise event predictions.

Results and Discussion

Here we show that urban crime may be predicted reliably one or more weeks in advance, enabling model-based simulations that reveal both the pattern of reported infractions and the pattern of corresponding police enforcement. We learn from recorded historical event logs, and validate on events in the following year beyond those in the training sample. Using incidence data from the City of Chicago, our novel spatio-temporal network inference algorithm infers patterns of past event occurrences, and constructs a communicating network (the Granger Network) of local estimators to predict future infractions. In this study, we consider two broad categories of reported criminal infractions: *violent crimes* consisting of homicides, assault, and battery, and *property crimes* consisting of burglary, theft and motor-vehicle thefts. The number of individuals arrested during each recorded event is separately modeled and allows us to investigate the possibility and pattern of enforcement bias.

We begin by processing event logs to obtain time-series of relevant events, stratified by location and discretized by time, yielding sequential event streams for 1) violent crime (v), 2) property crime (u) and 3) number of arrests (w), as shown in Fig. 1, panels a, b and c. To infer the structure of the Granger Net, we learn a finite state probabilistic transducer[15,16] for each possible source-target pair s, r and time lag Δ (Fig. 1d), yielding ≈ 2.6 billion modeled associations. Following the notion of Granger causality[17], links in the network are retained as they predict events at the target better than the target can predict itself. More details on the on problem characteristics and performance are provided in Tab. I and II respectively.

For Chicago, we make predictions separately for violent and property crimes, individually within spatial tiles roughly $1000ft$ across and time windows of 1 day approximately a week in advance with AUCs ranging from $80-99\%$ across the city. We summarize our prediction results in Fig. 2, where panels a and b illustrate the geospatial scatter of AUC obtained for different spatial tiles and types of crime, and c shows the distribution of AUCs achieved. Out-of-sample predictive performance remains stable over time; our predictions on successive years (each time using three preceding years for training, and one year for out-of-sample test, see Fig. 8 shows little variation in average AUC. Inspecting excerpts of the average daily crime rate for successive years also shows close match between actual and predicted behavior (See Fig. 9, panels a, c and e.) The remaining panels (b, d and f) in the same figure illustrate how the Fourier coefficients match up, showing that we are able to capture periodicities at the weekly and bi-weekly scales, and beyond.

Unlike previous efforts[1–5], we do not impose pre-defined spatial constraints. In contrast to contiguous diffusion phenomena encountered in physical systems, crime may spread across the complex landscape of a modern city unevenly, with regions hyperlinked by transportation networks, socio-demographic similarity, or historical collocation. Rather than assuming that events far off across the city will have a weaker influence compared with those physically near in space or time, we probe the topological structure emergent in the inferred dependencies to estimate the shape, size and organization of neighborhoods that best predict events at each location. The results illustrated in Fig. 2d and e show that the situation is complex with the locally predictive neighborhoods varying widely in geometry and size, implying that restricting analysis to relatively small local communities within the city is sub-optimal for crime prediction and enforcement analysis. In order to analyze if the effect of reported criminal infractions diffuse outward in space and time, we simply calculate temporal-spatial distances of influences, then average across all neighborhoods in the city, revealing the rapid decay with time delay in diffusion rates shown in Fig. 2f. Interestingly we find the property and violent crimes differ in their rates of influence diffusion (Fig. 2f); while the effect of property crimes decays rapidly in days, violent reported events shape the dynamics for weeks to come.

Forecasting crime via analyzing historical patterns has been attempted before[18,19]. These approaches use state of the art machine deep learning tools based on recurrent and convolutional neural networks (NN). In the first article[18], the authors train a NN model to predict next-day events for $60,348$ sample points in Chicago. The

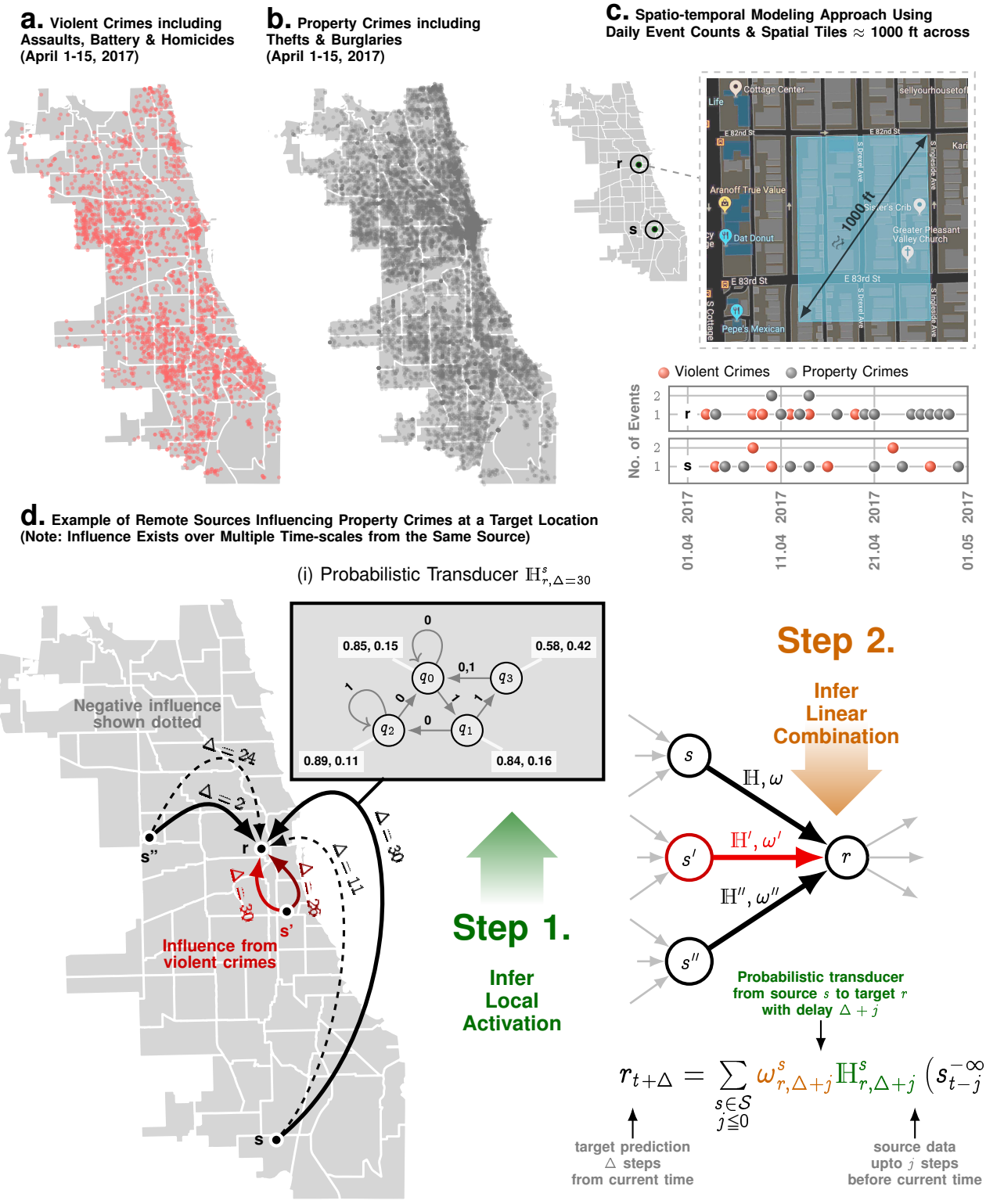

Fig. 1. **Crime Data & Modeling Approach**. **a** and **b** show the recorded infractions within the 2 week period between April 1 and 15 in 2017. Plate c illustrates our modeling approach: We break city into small spatial tiles approximately 1.5 times the size of an average city block, and compute models that capture multi-scale dependencies between the sequential event streams recorded at distinct tiles. In this paper, we treat violent and property crimes separately, and show that these categories have intriguing cross-dependencies. Plate d illustrates our modeling approach. For example, to predict property crimes at some spatial tile r, we proceed as follows: Step 1) we infer the probabilistic transducers that estimate event sequence at r by using as input the sequences of recorded infractions (of different categories) at potentially all remote locations (s, s', s'' shown), where this predictive influence might transpire over different time delays (a few shown on the edges between s and r). Step 2) Combine these weak estimators linearly to minimize zero-one loss. The inferred transducers can be thought of as inferred local activation rules, which are then linearly composed, reversing the approach of linearly combining input and then passing through fixed activation functions in standard neural net architectures. The connected network of nodes (variables) with probabilistic transducers on the edges comprises the Granger Network.

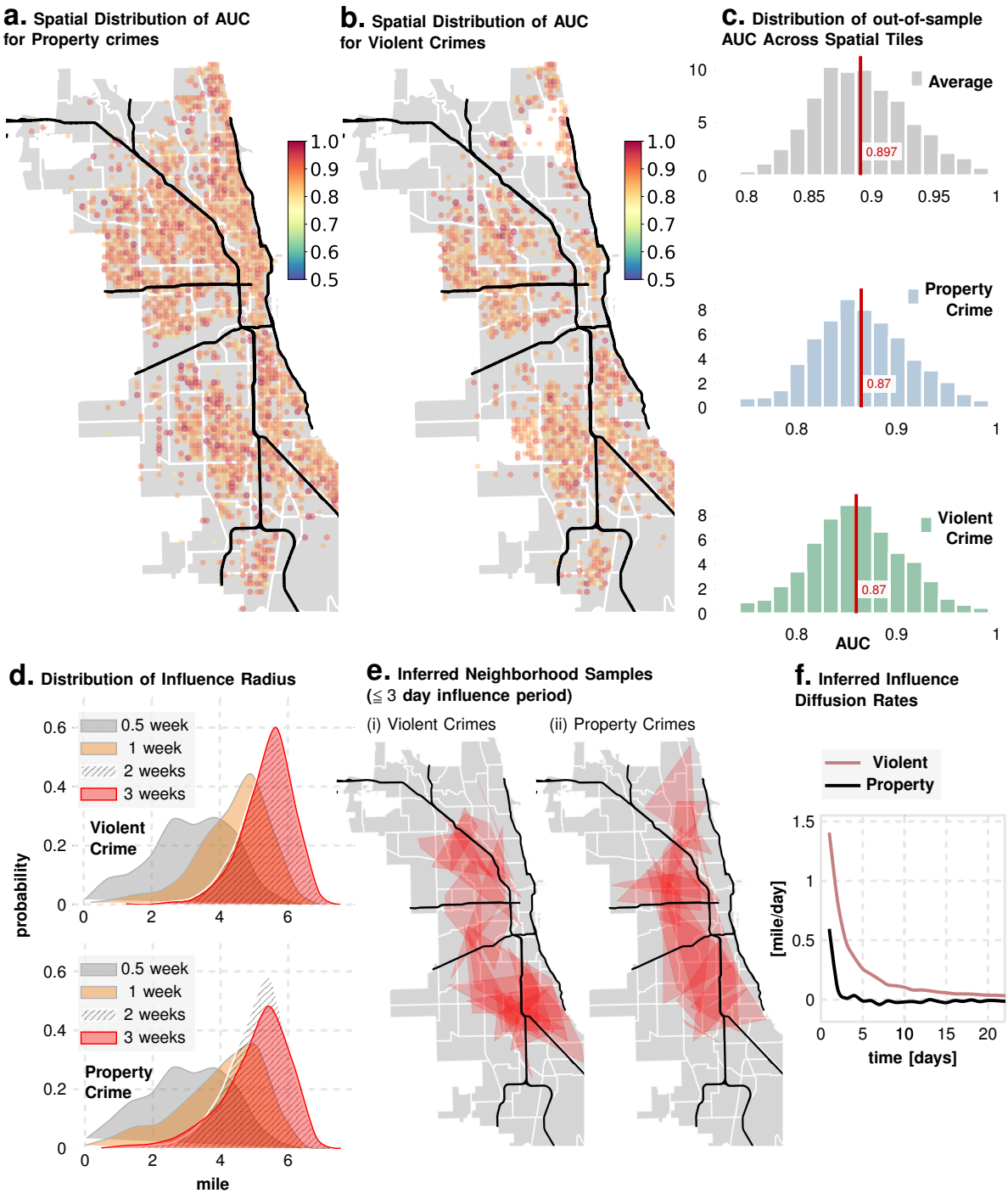

Fig. 2. **Predictive Performance of Granger Nets. a** an **b** illustrate the out-of-sample area under the receiver operating characteristics curve (AUC) for predicting violent and property crimes respectively. The prediction is made a week in advance, and the event is registered as a successful prediction if we get a hit within ± 1 day of the predicted date. **c** illustrates the distribution of AUC on average, individually for violent and property crimes. Our mean AUC is close to 90%. **Panels d-f shows influence Diffusion & Perturbation Space.** If we are able to infer a model that is predicts event dynamics at a specific spatial tile (the target) using observations from a source tile Δ days in future, then we say the source tile is within the influencing neighborhood for the target location with a delay of Δ. **d** illustrates the spatial radius of influence for 0.5, 1, 2 and 3 weeks, for violent (upper panel) and property crimes (lower panel). Note that the influencing neighborhoods, as defined by our model, are large and approach a radius of 6 miles. Given the geometry of the City of Chicago, this maps to a substantial percentage of the total area of urban space under consideration, demonstrating that crime manifests demonstrable long-range and almost city-wide influence. **e** illustrates the extent of a few inferred neighborhoods at time delay of at most 3 days. **f** illustrates the average rate of influence diffusion measured by number of predictive models inferred that transduce influence as we consider longer and longer time delays. Note that the rate of influence diffusion falls rapidly for property crimes, dropping to zero in about a week, whereas for violent crimes, the influence continues to diffuse even after three weeks.

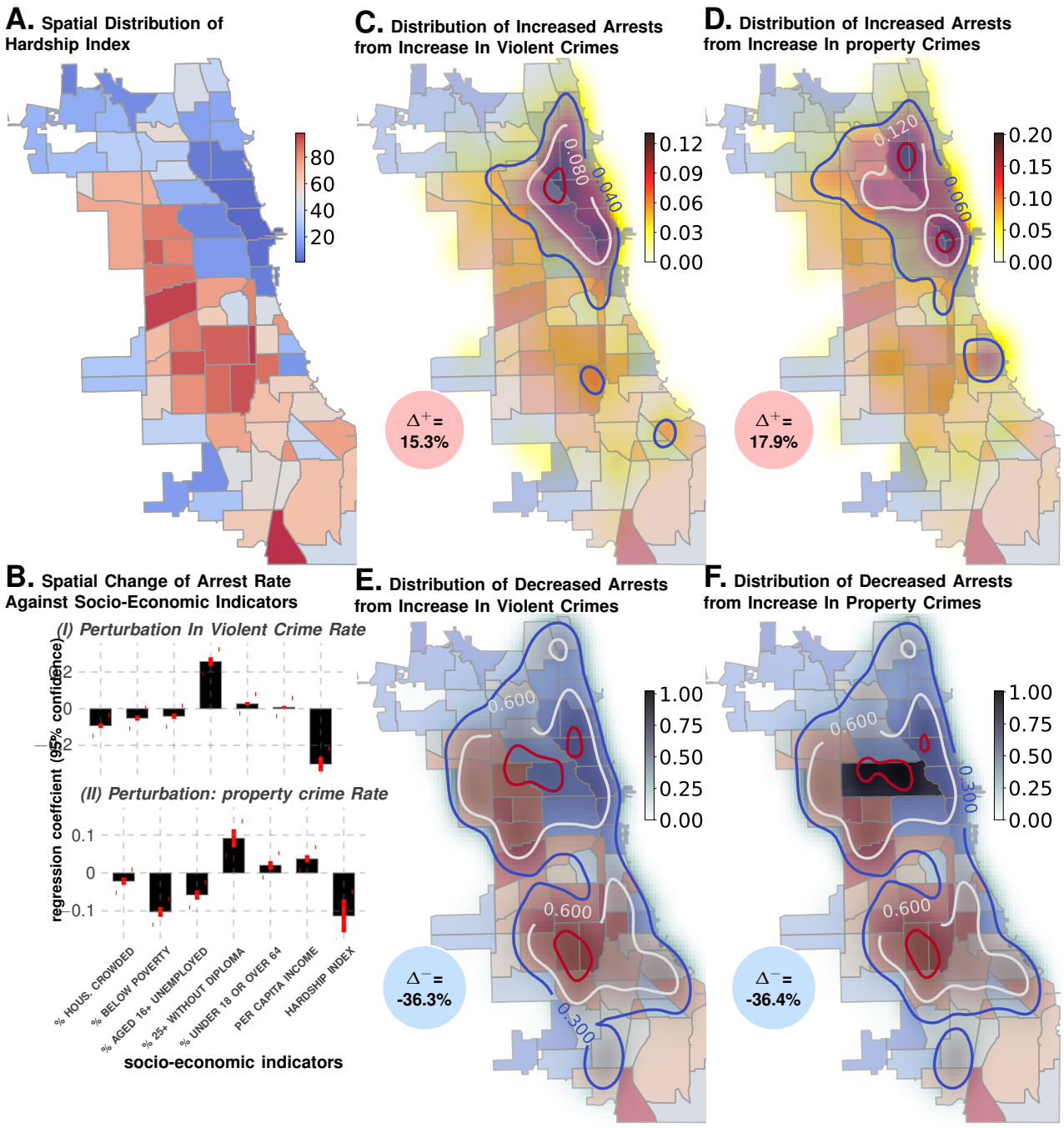

Fig. 3. **Estimating Bias. a** illustrates the distribution of hardship index (see SI). **c**, **d**, **e**, and **f** suggest biased response to perturbations in crime rates. With a 10% increase in violent or property crime rates, we see an approximately a 30% decrease in arrests when averaged over the city. The spatial distribution of locations that experience a positive vs. negative change in arrest rate reveals a strong preference favoring wealthy locations. If neighborhoods are doing better socio-economically, increased crime predicts increased arrests. A strong converse trend is observed in predictions for poor and disadvantaged neighborhoods, suggesting that under stress, wealthier neighborhoods drain resources from their disadvantaged counterparts. **b** illustrates this more directly via a multi-variable regression, where hardship index is seen to make a strong negative contribution.

model is trained on crime statistics, demographic makeup, meteorological data, and Google street view images to track graffiti, achieving an out-of-sample AUC of 83.3%. Our AUC is demonstrably higher (see Table II), and we predict with significantly less data (only past events), and 7 days into future (instead of next-day). Additionally, the use of demographic and graffiti is problematic with the possibility of introducing racial and socio-economic bias, with dubious causal value. In the second article[19], the authors combine convolutional and recurrent neural networks with weather, socio-economic, transportation, and crime data, to predict the next-day count of crime in Chicago. As spatial tiles, the authors use standard police beats, which break up Chicago into 274 regions. Police beats reflect the classical notion of social neighborhoods, and measure approximately 1 sq. mile on average[20]. In comparison, our spatial times are approximately 0.04 sq. miles, representing a 2500% higher resolution. This

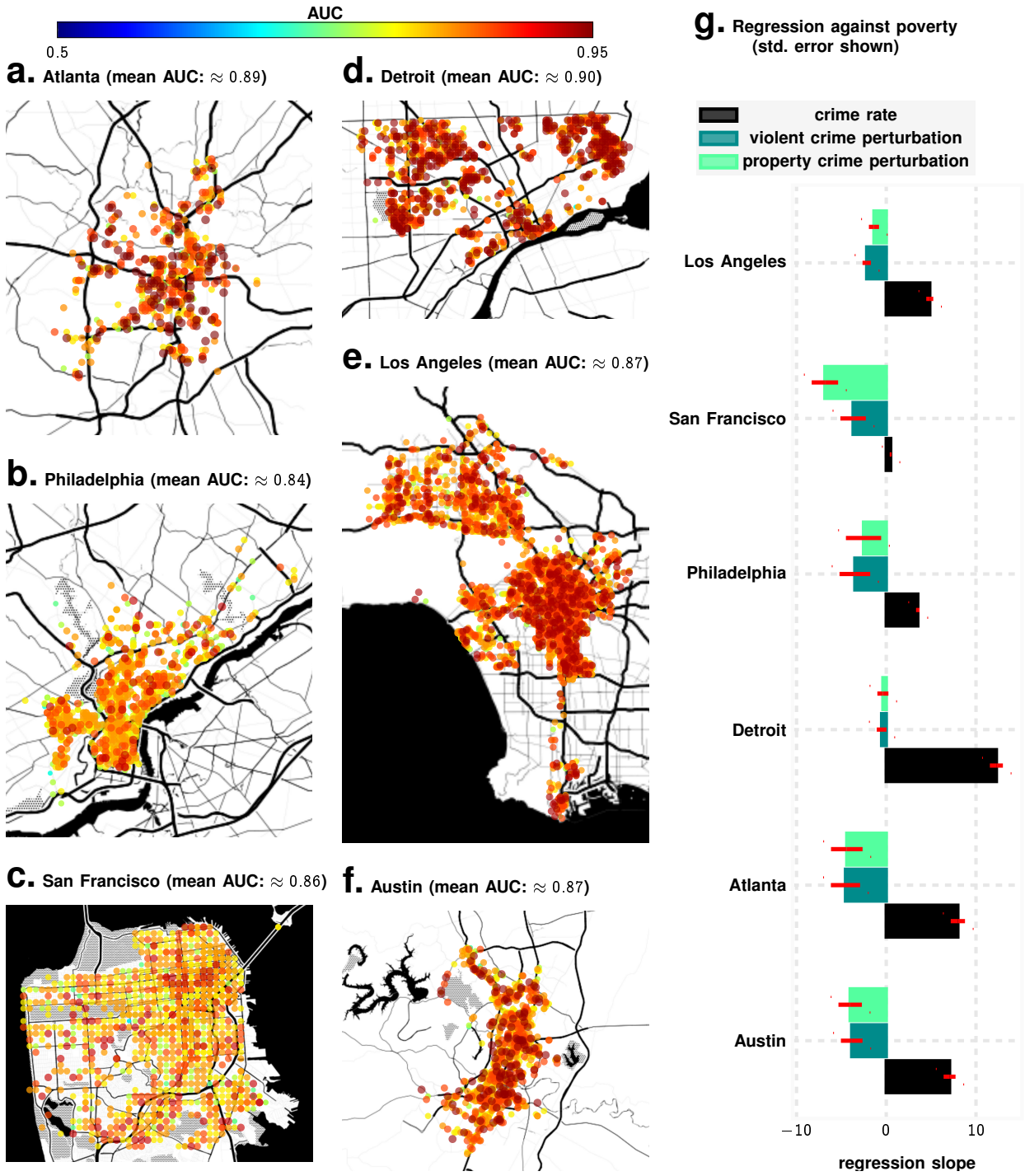

Fig. 4. **Prediction of property and violent crimes across major US cities and dependence of perturbation response on socio-economic status of local neighborhoods.** Panels a-f illustrate the AUCs achieved in six major US cities. These cities were chosen on the basis of the availability of detailed event logs in the public domain. All of these cities show comparably high predictive performance. Panel g illustrates the results obtained by regressing crime rate and perturbation response against SES variables (shown here for poverty, as estimated by the 2018 US census). We note that while crime rate typically goes up with increasing poverty, the number of events observed one week after a positive perturbation of 5-10% increase in crime rate is predicted to fall with increasing poverty. We suggest that this decrease is explainable by reallocation of enforcement resources disproportionately, away from disadvantaged neighborhoods in response to increased event rates, which leads to smaller number of reported crimes.

model achieves a classification accuracy of 75.6% for Chicago, which compares against our accuracy of $> 90\%$ (See Table II). While this competing model tracks more crime categories, it is limited to next-day predictions with significantly coarser spatial resolution. We also compare the predictive ability of naive autoregressive baseline models (See Material and Methods and Table III), which perform poorly, but provide a yardstick to meaningfully compare our claimed performance estimates.

With our precise predictive apparatus in place, we run a series of computational experiments that perturb the rates of violent and property crimes, and log the resulting alterations in future event rates across the city. By inspecting the effect of socio-economic status (SES) on the perturbation response, we investigate whether enforcement and policy biases modulate outcomes. The inferred stress response of the city suggests the presence of socio-economic bias (See Fig. 3). Wealthier neighborhoods away from the inner city respond to elevated crime rates with increased arrests, while arrest rates in disadvantaged neighborhoods drop, but the converse does not occur (See Fig. 3, panels e and f). Resource constraints on law enforcement, combined with biased prioritization to wealthier neighborhoods, result in reduced enforcement across the remainder of the city. This provides evidence for enforcement bias within U.S. cities that parallels widely discussed notions of suburban bias in wealthy suburbs[21,22]. While self-evident at the scale of countries and regions, the existence of unequal resource allocation in cities, where political power and influence concentrates in selective, wealthy neighborhoods, has been widely suspected[23–27]. Our analysis provides direct support for this contention, which shows up robustly for all years analyzed, going back over one and half decades in Chicago. Figs 6 and 7 show that these patterns are stable over time, at least in recent years. Additionally, Fig. 5 show the effect of perturbations across all variables, suggesting that crime reduction from perturbations seems to be most effective in regions with high crime rates, with SES confounders.

Beyond Chicago, we analyze criminal event logs available in the public domain for six additional major US cities: Detroit, Philadelphia, Atlanta, Austin, San Francisco and Los Angeles. In all these cities we obtain comparably high performance in predicting violent and property crimes, with average AUC ranging between 86-90% (See Fig 4a-f). In addition, our observed pattern of perturbation responses in Chicago, which suggests de-allocation of policing resources from disadvantaged neighborhoods to advantaged ones, is replicated in all these cities. While crime rate increases with degrading SES status of local neighborhoods, the number of predicted events a week after a positive 5-10% increase in crime rate is predicted to go down. Thus increasing the crime rate leads to a smaller number of reported crimes, a pattern holding more often in poorer neighborhoods.

Our analysis also sheds light on the continuing debate over the choice for neighborhood boundaries in urban crime modeling[28–31]. In Fig. 2d-f, we demonstrate that despite apparent natural boundaries, influence is often communicated over large distances and decays slowly, especially for violent crimes. More importantly, this study reveals how the “correct” choice of spatial scale should not be a major issue in sophisticated learning algorithms where optimal scales can be inferred automatically. We find that there exists a skeleton set of spatial tiles, which have strong influence on the overall event patterns (See Fig. 10). These induce a cellular decomposition of the city that identifies functional neighborhoods, where the cell-size adapts automatically to the local event dynamics.

LIMITATIONS & CONCLUSION

To our knowledge, this is the first analysis exploring perturbations of predictive data-driven models to probe the social dynamics of crime and its enforcement. Our ability to probe for the extent of enforcement bias is limited by our dataset; since inference of crime patterns are easily skewed by arrest rates. Disproportionate police response in Black communities can contribute to biases in event logs, which might propagate into inferred models. This has resulted in significant pushback from diverse communities against predictive policing[32]. Our approach is free from manual encoding of features (and thus resistant to implicit biases of the modelers themselves), but biases arising from disproportionate surveillance might still remain.

Even with its current limitations, however, this new addition to the toolbox of computational social science enables validation of complex theory from observed event incidence, supplementing the use of measurable proxies and potential biases in questionnaire-based data collection strategies. While classical approaches[33–36] broaden our understanding of the societal forces shaping both urban and regional landscapes, these approaches have neither successfully attempted to forecast individual infraction reports, nor reveal how these predictive patterns manifest systematic enforcement bias. In this study, we show how the ability of Granger Networks to predict such events not only opens new doors for precise intervention, but also advances the diagnosis and explanation of complex social patterns. We acknowledge the danger that powerful predictive tools place in the hands of over-zealous states in the name of civilian protection, but here we demonstrate their unprecedented ability to audit enforcement biases and hold states accountable in ways inconceivable in the past. We encourage widespread debate regarding how these technologies are used to augment state action in public life, and call for transparency that allows for continuous evaluation, reconsideration and critique.

MATERIALS AND METHODS

In this study we use historical geolocated incidence data of criminal infractions to model and predict future events in Chicago, Philadelphia, San Francisco, Austin, Los Angeles, Detroit and Atlanta. Each of the cities considered have a specific temporal and spatial resolution, which are optimized to maximize predictive performance (See Table I). The predictive performance obtained in these cities are enumerated in Table II. The distribution of AUCs obtained in Chicago for earlier years (2014-2017, predicted individually) are shown in Fig 8.

Data Source

The sources of the crime incidence data used in this study for the different US cities are enumerated in Table I. Theses logs include spatio-temporal event localization along with the nature, category, and a brief description of the recorded incident. For the City of Chicago, we also have access to the number of arrests made during or as a result of each event. For Chicago, the log is updated daily, keeping current with a lag of 7 days, and we make predictions for each of the years 2014-2017 (using 3 years before the target year for model inference, and 1 year for out-of-sample validation) for the prediction results shown in Figure 1. The evolving nature of the urban scenescape[37] necessitates that we restrict the modeling window to a few years at a time. The length of this window is decided by trading off loss of performance from shorter data streams to that the evolution of underlying generative processes for longer streams. The training and testing periods of the other cities is tabulated in Table I. In this study, we consider two broad categories of criminal infractions: *violent crimes* consisting of homicides, assault, battery etc., and *property crimes* consisting of burglary, theft, motor vehicle theft etc. Drug crimes are excluded from our consideration due to the possibility of ambiguity in the use of violence in such events. For the City of Chicago, the number of individuals arrested during each recorded event is considered a separate variable to be modeled and predicted, which allows us to investigate the possibility of enforcement biases in subsequent perturbation analyses.

We also use data on socio-economic variables available at the portal corresponding to Chicago community areas and census tracts, including % of population living in crowded housing, those residing below the poverty line, those unemployed at various age groups, per capita income, and the urban hardship index[38]. Such data is also obtained from the City of Chicago data portal. Additionally, we use data on poverty estimates for the other cities, which are obtained https://www.census.gov.

Spatial and Temporal Discretization & Event Quantization

Event logs are processed to obtain time-series of relevant events, stratified by occurrence locations. This is accomplished by choosing a spatial discretization, and focusing on one individual spatial tile at a time, which allows us to represent the event log as a collection of sequential event streams (See Fig. 1c). Additionally, we discretize time, and consider the sum total of events recorded within each time window.

Coarseness of these discretizations reflects a trade-off between computational complexity and event localization in space and time. Spatial and temporal discretizations are not independently chosen; a finer spatial discretization dictates a coarser temporal quantization, and vice versa to prevent long no-event stretches and long periods of contiguous event records, both of which reduce our ability to obtain reliable predictors. For the City of Chicago, we fix the temporal quantization to 1 day, and choose a spatial quantization such that we have high empirical entropy rates for the time series obtained. This results in spatial tiles measuring $0.00276° \times 0.0035°$ in latitude and longitude respectively, which is approximately $1000'$ across, roughly corresponding to an area of under 2×2 city blocks. Thus, any two points within our spatial tile are at worst in neighboring city blocks. We dropped from our analysis the tiles that have too low a crime rate ($< 5\%$ of days within the modeling window had any event recorded) to reduce computational complexity, resulting in an $N = 2205$ of spatial tiles in the city of Chicago.

The temporal and spatial resolution is adjusted in a similar manner for the other cities (See Table I).

Thus, we end up with three different integer-valued time series at each spatial tile: 1) violent crime (v), 2) property crime (u) and 3) number of arrests (w) in the City of Chicago. For other cities, we have only the first two categories, since information on arrests was not available. We ignore the magnitude of the observations, and treat them as Boolean variables. Thus, our models simply predict the presence or absence of a particular event type in a discrete spatial tile within a neighboring city block and observation window, *i.e.*, within the temporal resolution chosen, which is 1 day except for Atlanta, where is it is chosen to be 2 days (See Table I).

Inferring Generators of Spatio-temporal Cross-dependence

Let $\mathcal{L} = \{\ell_1, \cdots, \ell_N\}$ be the set of spatial tiles, and $\mathcal{E} = \{u, v, w\}$ be the set of event categories as described in the last section. At location $\ell \in \mathcal{L}$ for variable $e \in \mathcal{E}$, at time t, we have $(\ell, e)_t \in \{0, 1\}$, with 1 indicating the presence of at least one event. The set of all such combined variables (space + event type) is denoted as $\mathcal{S}$, *i.e.*, $\mathcal{S} = \mathcal{L} \times \mathcal{E}$. Let $T = \{0, \cdots, M-1\}$ denote the training period consisting of M time steps. Because for any time t, $(\ell, e)_t$ is a random variable, our goal here is to learn its dependency relationships with its own past, and with other variables in S to accurately estimate its future distribution for $t > T$.

To infer the structure of our predictive model, we learn a finite state probabilistic transducer[16] (referred to as a Crossed Probabilistic Finite State Automata or a XPFSA[15]) for each possible source-target pair $s, r \in \mathcal{S}$. Given a sequence of events at the source, these inferred transducers estimate the distribution of events at target r for some future point in time. Ability to estimate such a non-trivial distribution indicates the presence of causal influence. Here we assume that causal influence from the source to the target manifests as the source being able to predict events occurring at the target, better than the target can do by itself. This interpretation follows from Granger's eponymous approach to statistical causality[39]. Importantly, we do not assume that the underlying processes are iid, or that the model has any particular linear structure. Additionally, such influence is not restricted to be instantaneous. The source events might impact the target with a time delay, *i.e.*, a specific model between the source and target might predict events delayed by an *a priori* determined number of steps $\Delta_{max} \geqq \Delta \geqq 0$ specific to the model. Here we model the influence structure for each integer-valued delay separately. Thus, for source s and target t, we can have $\Delta_{max} + 1$ transducers each modeling the influence for a specific delay in $\{0, \Delta_{max}\}$. The maximum number of steps in time delay Δ_{max} is chosen a priori, based on the problem at hand.

While these influences or dependencies may differ for different delays, they need not be symmetric between source and target pairs. The complete set, comprising at most $|\mathcal{S}|^2(\Delta_{max} + 1)$ models, represents a predictive framework for asymmetric multi-scale spatio-temporal phenomena. Note that the number of possible models increase quickly. For example, for the City of Chicago, for $\Delta_{max} = 60$ with 2205 spatial tiles and three event categories, the number of inferred models is bounded above by ≈ 2.6 billion.

ur approach consists of inferring XPFSAs in two key steps (See Fig. 1d, and discussion later in SI-Section 2): First, we infer XPFSA models for all source-target pairs and all delays up to Δ_{max}. In the second step, we learn a linear combination of these transducers to maximize predictive performance. Denoting the observed event sequence in time interval $(\infty, t]$ at source s as $s_t^{-\infty}$, the XPFSA $\mathbb{H}^s_{r,k}$ estimates the distribution of events for target r at time step $t + k$. This is accomplished by learning an equivalence relation on the historical event sequences observed at source s, such that equivalent histories induce an approximately identical future event distribution at target r, k steps in the future. Thus, for example, the XPFSA shown in Fig. 1d has four states, indicating that there are 4 such equivalence classes of observations that induce the distinct output probabilities shown from each state. Often this estimate is not very precise due to the possibility for multi-scale and multi-source influence, *e.g.*, when target r is influenced by multiple sources with different time delays. In the second step, we employ a standard gradient boosting regressor for each target, to optimize the linear combination of inferred transducers and learn the scalar weights $\omega^s_{r,k}$ for source s, target r and delay k. Detailed pseudocode of the inference algorithms are provided in the SI-section 1.

To compare with a standard neural net architecture, these probabilistic transducers may be viewed as local non-linear activation functions. With neural networks we repeatedly compute affine combination of inputs and apply fixed non-linear activation to the combined input and finally optimize affine combination weights via backpropagation, but here we first learn the local non-linear activations, and then optimize the linear or affine combination of weak estimators. Optimizing the weights is a significantly simpler, local operation and may be done with any standard regressor. In contrast to recurrent neural nets (RNN), the role of hidden layer neurons is partially accounted for by states of the XPFSA, which are a priori undetermined both with respect to their multiplicity and their transition connectivity structure.

Computational & Model Complexity

We assume the maximum time delay in the influence propagation to be 60 days for all cities, which for the City of Chicago results in at most $2,669,251,725$ inferred models, of which $61,650,000$ are useful with $\gamma \geqq 0.01$. Model inference in this case consumed approximately $200K$ core-hours on 28 core Intel Broadwell processors, when carried out with incidence data over the period Jan 1, 2014 to December 31, 2016. Computational cost for other time-periods and other cities are comparable and roughly scale with the square of the number of spatial tiles, and linearly with the length of time-quantized data-streams considered as input to the inference algorithm.

Crime Prediction Metrics

For each spatial location, the inferred Granger Net maps event histories to a raw risk score as a function of time. The higher this value, the higher the probability of an event of target type occurring at that location, within the specified time window. To make crisp predictions, however, we must choose a decision threshold for this raw score. Conceptually identical to the notion of Type 1 and Type 2 errors in classical statistical analyses, the choice of a threshold trades off false positives (Type 1 error) for false negatives (Type 2 error). Choosing a small threshold results in predicting a larger fraction of future events correctly, *i.e.*, have a high true positive rate (TPR), while simultaneously suffering from a higher false positive rate (FPR), and vice versa. The receiver operating characteristic curve (ROC) is the plot of the FPR vs the TPR, as we vary this decision threshold. If our predictor is good, we will consistently achieve high TPR with small FPR resulting in a large area under the ROC curve denoted as the AUC. Importantly, AUC measures intrinsic performance, independent of the threshold choice. Thus, the AUC is immune to class imbalance (the fact that crimes are by and large rare events). An AUC of 50% indicates that the predictor does no better than random, and an AUC of 100% implies that we can achieve perfect prediction of future events, with zero false positives.

We use a flexible approach in evaluating AUC; a positive prediction is treated as correct if there is at least one event recorded in ± 1 time steps in the target spatial tile.

Predictability Analysis

In the City of Chicago, we can predict events approximately a week in advance at the spatial resolution of ± 1 city blocks with a temporal resolution of ± 1 day, with a false positive rate of less than 20% and a median true positive rate of 78%. The predictive performance in the other cities is enumerated in Table II. While not directly modeled in the frequency domain, we found that the event forecasts produce very similar signatures in the frequency domain (See Fig. 9), when compared over the first 150 days of each out-of-sample period (1 yr).

Spatial Neighborhoods

The degree of causal influence exerted by one variable (the source stream) on another (the target stream) is quantified by the coefficient of causal dependence (γ, see SI-Section 2). Identifying the source-target pairs for which the coefficient of causality is high (See Fig 10), we note that there exists a sparse set of spatial tiles which exert nearly all of the influence in the entire set of observed variables. Thus, observing these variables alone would enable us to make good event forecasts. These tiles span the expanse of the city, and a Voronoi decomposition based on the centers of these tiles in shown in Fig 10b. Such a decomposition demonstrates an algorithmic approach to choosing optimal neighborhoods for urban analysis.

Perturbation Analysis

We experimented with positive and negative perturbations to both violent and property crime rates ranging from 1 to 10% of observed rates. Response to perturbed crime rates was measured as the relative change from nominal baseline in estimated time-average for the predicted event frequencies 1 week in the future, corresponding to violent and property crimes and number of arrests.

Results from our perturbation experiments shed light both on the stability characteristics of crime in Chicago, and further allowed us to look for evidence of biased police enforcement responses under stress. Under stress, well-off neighborhoods tend to drain resources disproportionately from disadvantaged locales (See Fig. 3). For economically well-off neighborhoods in the bottom 25% of the hardship index are much more likely to see a near -proportional increase ($\approx 15\%$) in law enforcement response, measured by the number or predicted arrests on a 10% increase in crime rates (See Fig. 3, panels c and d, which show how regions with increased enforcement response are concentrated in well-off neighborhoods), while the rest of the city see a drop in predicted response of about twice the magnitude ($> 30\%$). Increased crimes causes enforcement resources to be drained from disadvantaged neighborhoods to support their better socioeconomic counterparts. We performed multi-variable linear regression analysis to evaluate the question in another way. Here we regressed violent and property crime rates, independently, on the variables listed in (Fig. 3b), including a slope intercept variable in each model. In both models, the hardship index exhibits a strong, negative influence on changes in arrest rate from perturbations that increase violent and the property crime rates, which contradicts what might be expected in the absence of bias. Poorer neighborhoods have more crime and so these socio-economic indicators should contribute positively to the arrest rate with increasing crime. These patterns were replicated in our perturbation experiments for all preceding years we analyzed (2014 through 2017, See Fig 6 and 7). Response measured in the property an violent crimes, and in the associated arrests from perturbations is detailed in Fig 5.

TABLE I
CRIME EVENT LOG INFORMATION FOR CITIES CONSIDERED

	Atlanta	Austin	Detroit	Los Angeles	Philadelphia	San Francisco	Chicago
no. of variables[1]	510	1082	1161	3287	1037	975	3826
temporal resolution	2 days	1 day	1 day	1 day	1 day	1 day	1 day
bounding box of modeled region	33.65°N, 33.86°N, 84.54°W, 84.31°W	30.14°N, 30.48°N, 97.89°W, 97.63°W	42.30°N, 42.45°N, 83.28°W, 82.91°W	33.71°N, 34.33°N, 118.65°W, 118.16°W	39.88°N, 40.12°N, 75.27°W, 74.96°W	37.71°N, 37.81°N, 122.51°W, 122.36°W	41.64°N, 42.06°N, 87.88°W, 87.52°W
spatial resolution	$983' \times 983'$	$983' \times 983'$	$983' \times 983'$	$983' \times 983'$	$983' \times 983'$	$983' \times 983'$	$951' \times 1006'$
Spatial exclusion threshold[2]	2.5%	2.5%	2.5%	2.5%	5.0%	2.5%	5.0%
training period	14/01/01-18/12/31	16/01/01-18/12/31	12/01/01-14/12/31	16/01/01-18/12/31	16/01/01-18/12/31	14/01/01-16/12/31	14/01/01-16/12/31
test period	19/01/01-19/07/20	19/01/01-19/04/11	15/01/01-15/04/11	19/01/01-19/04/11	19/01/01-19/04/11	17/01/01-17/04/11	17/01/01-17/04/11
prediction horizon	6 days	3 days	3 days	3 days	3 days	3 days	7 days
violent crime stat.	event count 2649, rate 3.98%	event count 20132, rate 5.45%	event count 20922, rate 3.72%	event count 72355, rate 4.83%	event count 33803, rate 8.11%	event count 23317, rate 7.16%	event count 179274, rate 7.7%
property crime stat.	event count 23522, rate 4.51%	event count 88929, rate 6.22%	event count 39840, rate 3.30%	event count 205435, rate 5.49%	event count 85683, rate 9.02%	event count 197835, rate 12.83%	event count 263661, rate 7.0%
data source	opendata.atlantapd.org	data.austintexas.gov	data.detroitmi.gov	data.lacity.org	www.opendata\philly.org	data.sfgov.org	data.cityofchicago.org

[1] No. of variables indicates the total number of time series considered for violent and property crimes.

[2] Tiles with less than threshold event-rate were excluded.

TABLE II
PREDICTION PERFORMANCE WITH GRANGER NET FOR SEVEN US CITIES

city	property crimes		violent crimes	
	median AUC	accuracy[†]	median AUC	accuracy
Atlanta	0.90	0.84	0.88	0.84
Austin	0.87	0.82	0.88	0.83
Detroit	0.90	0.86	0.89	0.84
Philadelphia	0.87	0.81	0.87	0.81
Los Angeles	0.84	0.83	0.84	0.83
San Francisco	0.86	0.80	0.86	0.81
Chicago	0.87	0.93	0.87	0.94

[†] Accuracy calculated with sensitivity$\times$frequency$+$specificity$\times(1-$frequency$)$.

We also carried out similar perturbation analyses for the other cities, and observed that with increasing poverty we have expected increase of observed crime rates, but an unexpected decrease in violent and property crimes after a 5-10% simulated uptick in either category of crimes (See Fig. 4).

Naive Baselines: Autoregressive Integrated Moving Average (ARIMA) Models

To explore the predictive ability of naive baseline models on our datasets, we consider four ARIMA[?] configurations with lag orders $p = 5$ and 10, numbers of differencing $d = 1$ and 2, and the window of moving average $q = 0$. Let y_t be the series we want to model and y'_t be y_t differenced by d times, the ARIMA(p, d, q) models series y'_t by

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (1)$$

TABLE III
NAIVE BASELINE RESULTS: MEAN AUC ACHIEVED WITH ARIMA MODELS

city	**ARIMA**$(5,1,0)$	**ARIMA**$(10,1,0)$	**ARIMA**$(5,2,0)$	**ARIMA**$(10,2,0)$
Atlanta	0.65	0.66	0.62	0.66
Austin	0.65	0.68	0.63	0.67
Detroit	0.59	0.62	0.57	0.61
Philadelphia	0.64	0.65	0.63	0.65
Los Angeles	0.64	0.67	0.61	0.66
San Francisco	0.68	0.70	0.66	0.69
Chicago	0.70	0.71	0.67	0.69

where $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$ are the coefficients to be fitted. In Eq. (1), y'_{t-k}s are the historical values of y'_t whose inclusion models the influence of past values on the current value (autoregression), and ε_{t-k}s are the white noise terms whose inclusion models the dependence of current value against current and previous (observed) white noise error terms or random shocks (moving average). Specifically, we use the following four models for the earthquake and the crime datasets

$$y_t^{(1)} = c + \phi_1 y'_{t-1} + \cdots + \phi_5 y'_{t-5} \tag{2}$$

$$y_t^{(1)} = c + \phi_1 y'_{t-1} + \cdots + \phi_5 y'_{t-10} \tag{3}$$

$$y_t^{(2)} = c + \phi_1 y'_{t-1} + \cdots + \phi_5 y'_{t-5} \tag{4}$$

$$y_t^{(2)} = c + \phi_1 y'_{t-1} + \cdots + \phi_5 y'_{t-10} \tag{5}$$

where $y_t^{(d)}$ is y_t different by d times ($y_t^{(1)} = y_t - y_{t-1}$ and $y_t^{(2)} = y_t - 2y_{t-1} + y_{t-2}$). For simple benckmarking, we apply the ARIMA model to each individual time series, which means the predictive model is trained without exogenous variables. For the implementation, we use the Python `statsmodels` package[?], and the result is shown in Tab. III. The inadequate performance of ARIMA may be due to 1) the use of a single data stream limits the ability of ARIMA to capture the interplay between co-evoluting processes, and 2) a pre-determined lag order fails to capture the possibly varying temporal memory of individual processes.

ACKNOWLEDGMENTS

Our work greatly benefited from discussion of everyone who participated in our workshop series on crime prediction at the Neubauer Collegium for culture and society[40], and with those with whom we had extended conversations to ground and refine our modeling approach.

Data was provided by the City of Chicago Data Portal at https://data.cityofchicago.org. The City of Chicago ("City") voluntarily provides the data on this website as a service to the public. The City makes no warranty, representation, or guaranty as to the content, accuracy, timeliness, or completeness of any of the data provided at this website (https://www.chicago.gov/city/en/narr/foia/data_disclaimer.html), and the authors of this study are solely responsible for the opinions and conclusions expressed in this study. Sources of the crime incidence data for the other cities are tabulated in the Supplementary text. Socio-ecomonic data for metropolitan areas was obtained from https://www.census.gov.

This work is funded in part by the Defense Sciences Office of the Defense Advanced Research Projects Agency projects HR00111890043/P00004 and W911NF2010302, and the Neubauer Collegium for Culture and Society through the Faculty Initiated Research Program 2017. The claims made in this study do not necessarily reflect the position or the policy of the sponsors, and no official endorsement should be inferred.

REFERENCES

[1] Bowers, K. J., Johnson, S. D. & Pease, K. Prospective hot-spotting: The future of crime mapping? *The British Journal of Criminology* **44**, 641–658 (2004).

[2] Chainey, S., Tompson, L. & Uhlig, S. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal* **21**, 4–28 (2008).

[3] Fielding, M. & Jones, V. 'disrupting the optimal forager': Predictive risk mapping and domestic burglary reduction in trafford, greater manchester. *International Journal of Police Science & Management* **14**, 30–41 (2012).

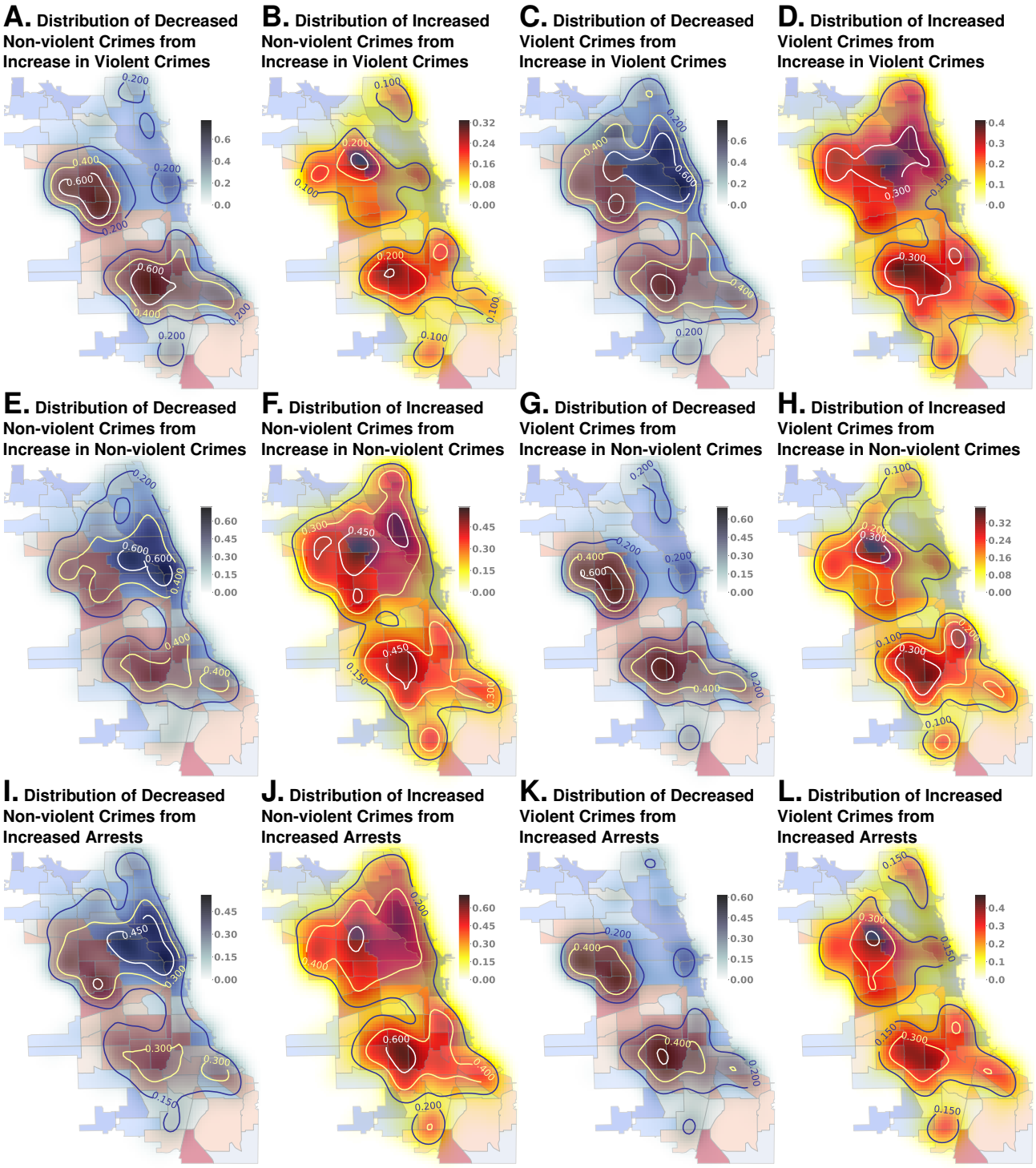

Fig. 5. **Perturbation Effects Across Variables.** We see that the decrease of **violent crimes from increase of property crimes** are localized in disadvantaged neighborhoods (panel g). Similarly, the decrease of **property crimes from increase of violent crimes** is also localized to disadvantaged neighborhoods (panel a), as well as the decreased violent crimes from increased arrests (panel k). We see a weaker localization for the corresponding increases in crime rates under similar perturbations. Looking at other pairs of variables under perturbation (rest of the panels), we generally do not see a very prominent correspondence with the distribution of socio-economic indicators. It seems crimes (and particulalrly violent crimes) are easier to dampen in lcales with high existing crime rates, which is desirable result. But such conclusions are currently confounded by SES variables, and futher work is needed to investigate these effects more thoroughly.

[4] Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P. & Tita, G. E. Self-exciting point process modeling of crime. *Journal of the American Statistical Association* **106**, 100–108 (2011).

[5] Mohler, G. O. *et al.* Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association* **110**, 1399–1411 (2015).

[6] Poisson, S. D. *Probabilité des jugements en matière criminelle et en matière civile, précédées des règles*

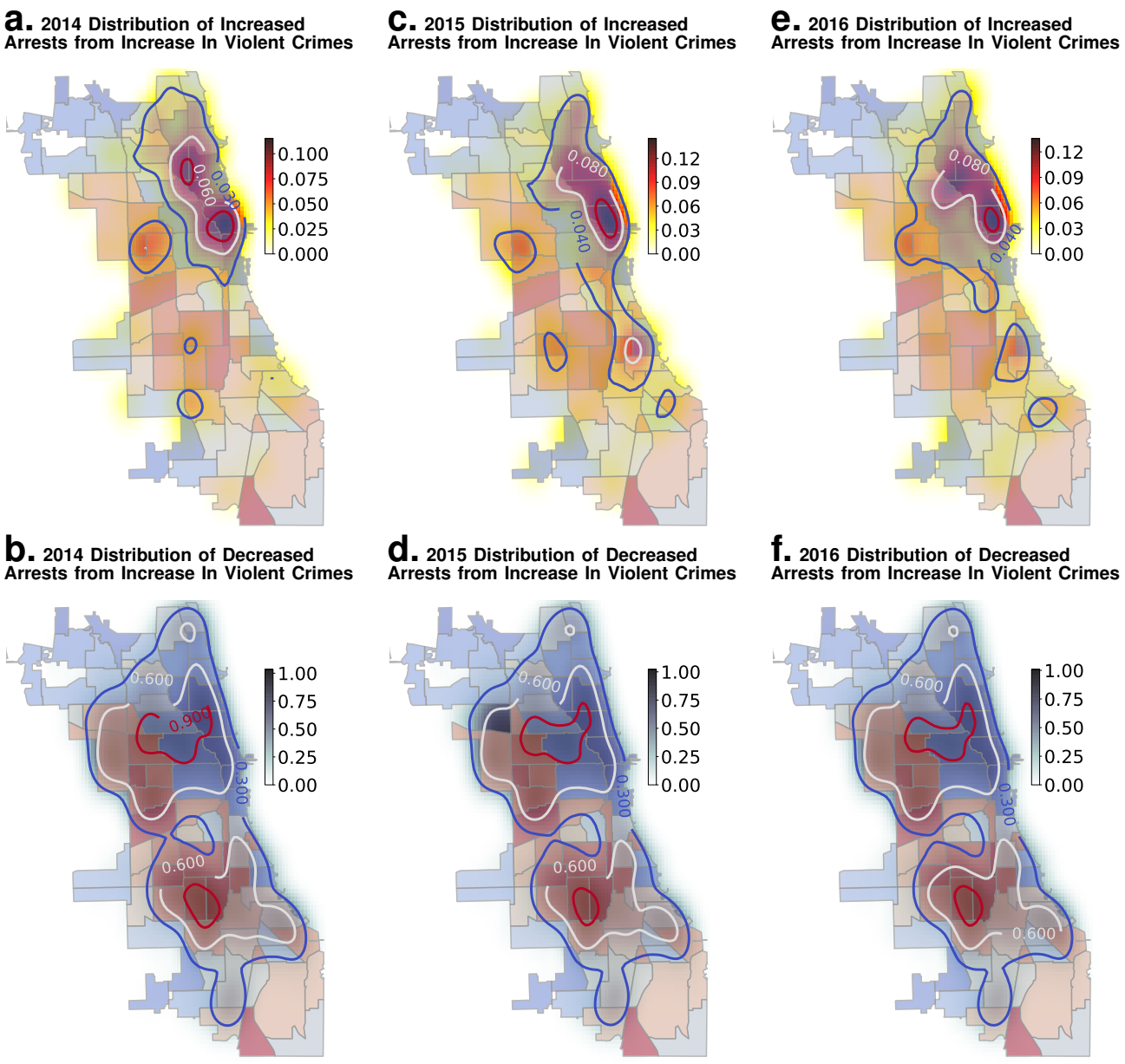

Fig. 6. **Stability of Suburban Bias over Years (Violent Crimes).** We show that the nature of the perturbation response shown in Fig. 3 in the main text holds true for earlier years as well: panels a and b correspond to year 2014, c and d correspond to 2015 and e and f correspond to year 2016, all of which follow the same pattern shown in Fig. 3 in the main text.

générales du calcul des probabilitiés (Bachelier, 1837).

[7] Du Sautoy, M. *The Creativity Code: Art and Innovation in the Age of AI* (Harvard University Press, 2020).

[8] Ferdinand, T. N. Demographic shifts and criminality: An inquiry. *The British Journal of Criminology* **10**, 169–175 (1970).

[9] Cohen, L. & Felson, M. Social change and crime rate trends: A routine activity approach. *American Sociological Review* **44**, 588–608 (1979). Cited By 4102.

[10] Cohen, L. E. Modeling crime trends: a criminal opportunity perspective. *Journal of Research in Crime and Delinquency* **18**, 138–164 (1981).

[11] Wang, X. & Brown, D. E. The spatio-temporal modeling for criminal incidents. *Security Informatics* **1**, 2 (2012).

[12] Liu, H. & Brown, D. E. Criminal incident prediction using a point-pattern-based density model. *International Journal of Forecasting* **19**, 603 – 622 (2003).

[13] Caplan, J. M., Kennedy, L. W., Barnum, J. D. & Piza, E. L. Crime in context: Utilizing risk terrain modeling and conjunctive analysis of case configurations to explore the dynamics of criminogenic behavior settings. *Journal of Contemporary Criminal Justice* **33**, 133–151 (2017).

[14] Kang, H. W. & Kang, H. B. Prediction of crime occurrence from multi-modal data using deep learning.

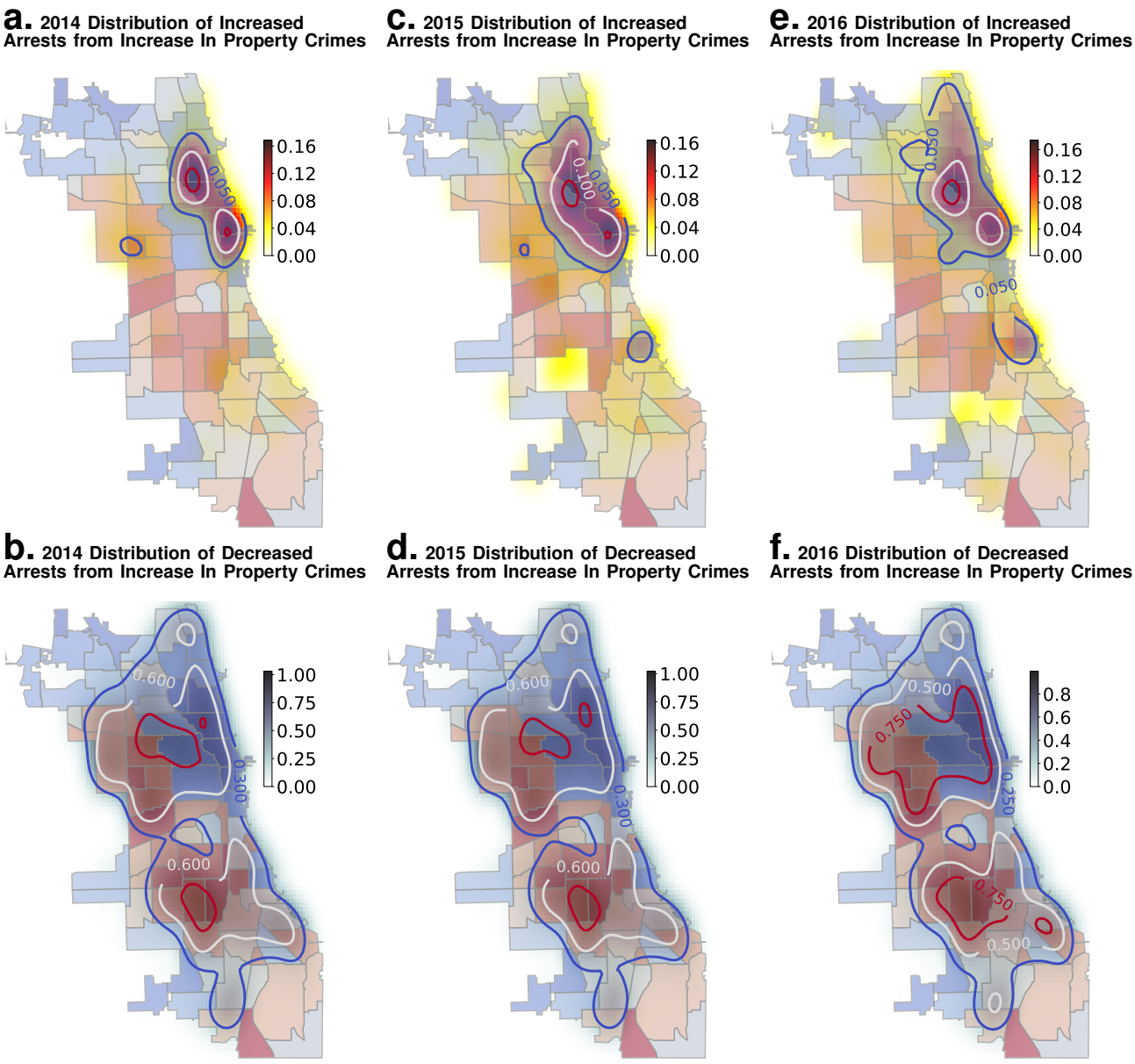

Fig. 7. **Stability of Suburban Bias over Years (Property Crimes).** We show that the nature of the perturbation response shown in Fig. 3 in the main text holds true for earlier years as well: panels a and b correspond to year 2014, c and d correspond to 2015 and e and f correspond to year 2016, all of which follow the same pattern shown in Fig. 3 in the main text.

PLoS ONE **12**, e0176244 (2017).

[15] Chattopadhyay, I. Causality networks. *arxiv CoRR* (2014). URL http://arxiv.org/abs/1406.6651.

[16] Mohri, M. *Weighted Finite-State Transducer Algorithms. An Overview*, 551–563 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2004).

[17] Granger, C. W. J. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control* **2**, 329 – 352 (1980).

[18] Kang, H.-W. & Kang, H.-B. Prediction of crime occurrence from multi-modal data using deep learning. *PloS one* **12**, e0176244 (2017).

[19] Stec, A. & Klabjan, D. Forecasting crime with deep learning. *arXiv preprint arXiv:1806.01486* (2018).

[20] Hannon, L. Neighborhood residence and assessments of racial profiling using census data. *Socius* **5**, 2378023118818746 (2019).

[21] Meyer, W. B. & Graybill, J. K. The suburban bias of american society? *Urban Geography* **37**, 863–882 (2016).

[22] Lipton, M. *et al. Why poor people stay poor: a study of urban bias in world development* (London: Canberra, ACT: Temple Smith; Australian National University Press, 1977).

[23] Jackson, K. T. *Crabgrass frontier: The suburbanization of the United States* (Oxford University Press, 1987).

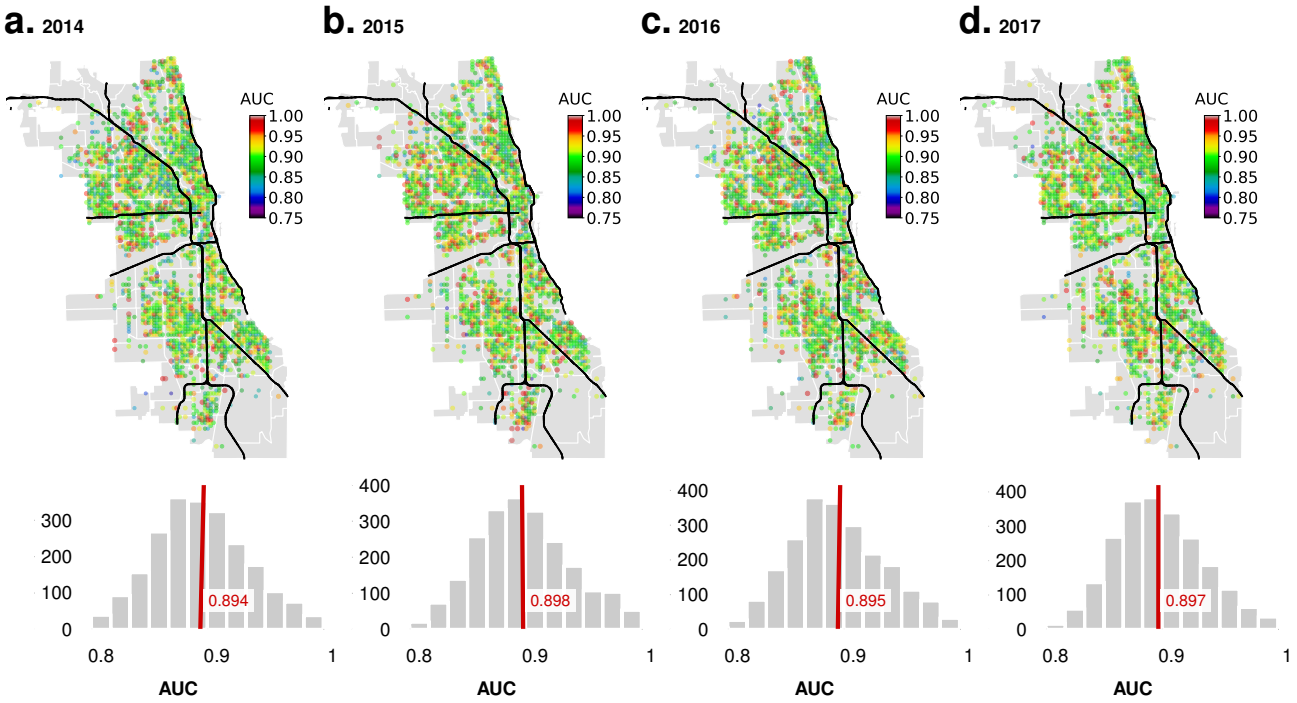

Fig. 8. **Out of Sample Predictive Performance over the Years.** We show that the predictive performance is very stable, and variation in mean AUC is limited to the third place of decimal, at least when analyzing the last few years (4 years shown).

[24] Duany, A., Plater-Zyberk, E. & Speck, J. *Suburban nation: The rise of sprawl and the decline of the American dream* (Macmillan, 2001).

[25] Logan, J. R. The suburban advantage: New census data show unyielding city-suburb economic gap, and surprising shifts in some places. *Lewis Mumford Center for Comparative Urban and Regional Research, University at Albany* (2002).

[26] Lazare, D. *America's Undeclared War: What's Killing Our Cities and how to Stop it* (Harcourt, 2001).

[27] Young, I. M. *Inclusion and democracy* (Oxford University press on demand, 2002).

[28] SHERMAN, L. W., GARTIN, P. R. & BUERGER, M. E. Hot spots of predatory crime: Routine activities and the criminology of place*. *Criminology* **27**, 27–56 (1989). https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-9125.1989.tb00862.x.

[29] WOOLDREDGE, J. Examining the (ir)relevance of aggregation bias for multilevel studies of neighborhoods and crime with an example comparing census tracts to official neighborhoods in cincinnati*. *Criminology* **40**, 681–710 (2002).

[30] MEARS, D. P. & BHATI, A. S. No community is an island: The effects of resource deprivation on urban violence in spatially and socially proximate communities*. *Criminology* **44**, 509–548 (2006). https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-9125.2006.00056.x.

[31] Weisburd, D., Groff, E. R., Yang, S.-M. & Telep, C. W. *Criminology of Place*, 848–857 (Springer New York, New York, NY, 2014).

[32] Predictive policing algorithms are racist. they need to be dismantled. — mit technology review. https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/. (Accessed on 01/29/2021).

[33] Sutherland, E. H. Juvenile delinquency and urban areas: A study of rates of delinquents in relation to differential characteristics of local communities in american cities. clifford r. shaw , henry d. mckay , norman s. hayner , paul g. cressey , clarence w. schroeder , t. earl sullenger , earl r. moses , calvin f. schmid. *American Journal of Sociology* **49**, 100–101 (1943). https://doi.org/10.1086/219339.

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

[35] Miethe, T. D., Hughes, M. & McDowall, D. Social Change and Crime Rates: An Evaluation of Alternative Theoretical Approaches*. *Social Forces* **70**, 165–185 (1991). http://oup.prod.sis.lan/sf/article-pdf/70/1/165/6887328/70-1-165.pdf.

[36] Braga, A. A. & Clarke, R. V. Explaining high-risk concentrations of crime in the city: Social disorganization, crime opportunities, and important next steps. *Journal of Research in Crime and Delinquency* **51**, 480–498

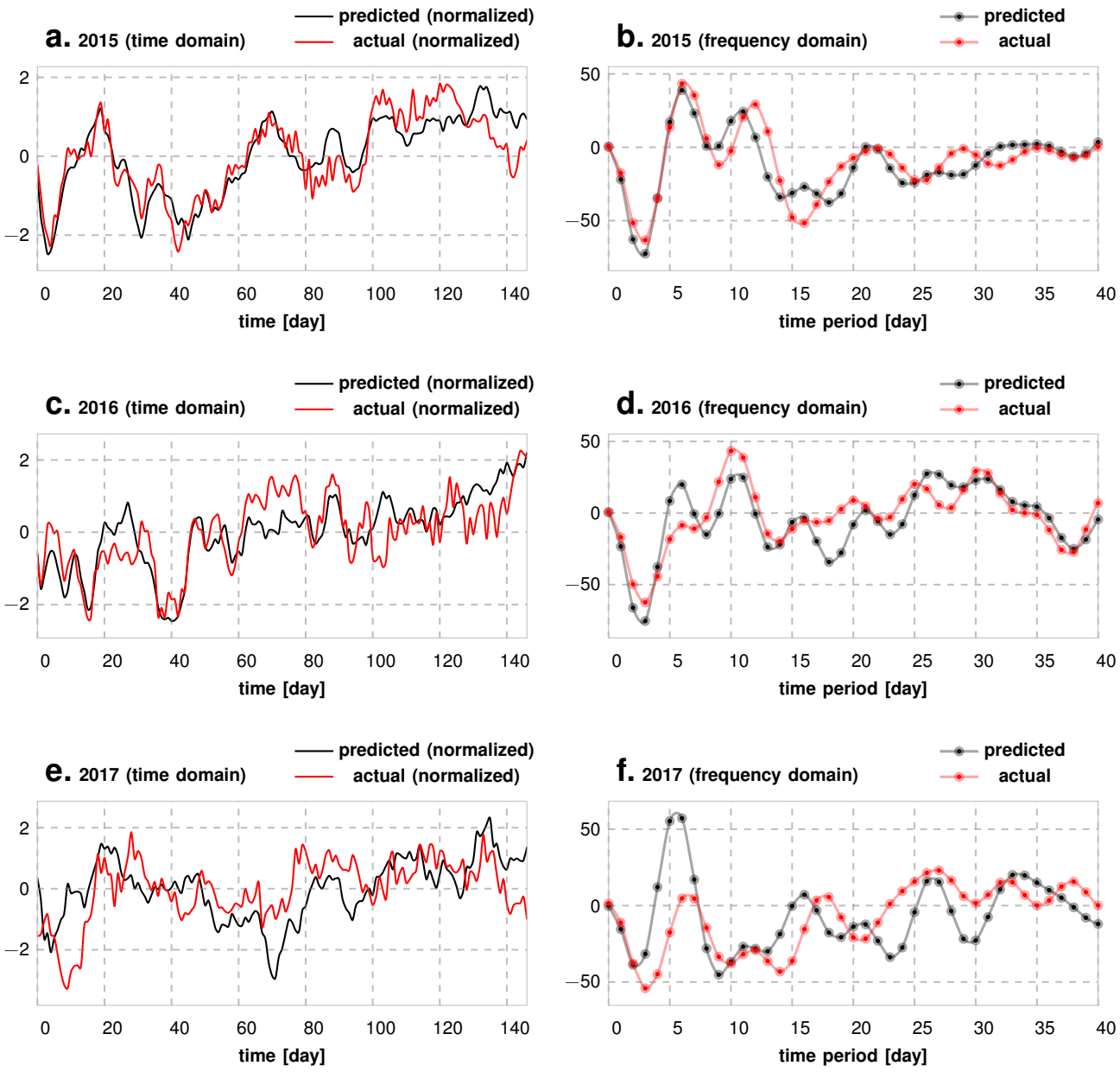

Fig. 9. **Comparison of Predicted vs Actual Sample Paths in Time and Frequency Domains.** Panels a, c and e show that the predicted and actual sample paths are pretty close for different years, when compared over the first 150 days of each year. Panels b, d and f show that the Fourier coefficients match up pretty well as well. More importantly, while our models do not explicitly incorporate any periodic elements that are being tuned, we still manage to capture the weekly, (approximately) biweekly and longer periodic regularities.

(2014).

[37] Silver, D. & Clark, T. *Scenescapes: How Qualities of Place Shape Social Life* (University of Chicago Press, 2016).

[38] Nathan, R. P. & Adams, C. F. Four perspectives on urban hardship. *Political Science Quarterly* **104**, 483–508 (1989).

[39] Granger, C. W. J. Testing For Causality. *Journal of Economic Dynamics and Control* **2**, 329–352 (1980).

[40] University of Chicago. Crimes of prediction workshop, the neubauer collegium for culture and society. https://neubauercollegium.uchicago.edu/events/uc/crimes_of_prediction_workshop/ (2019).

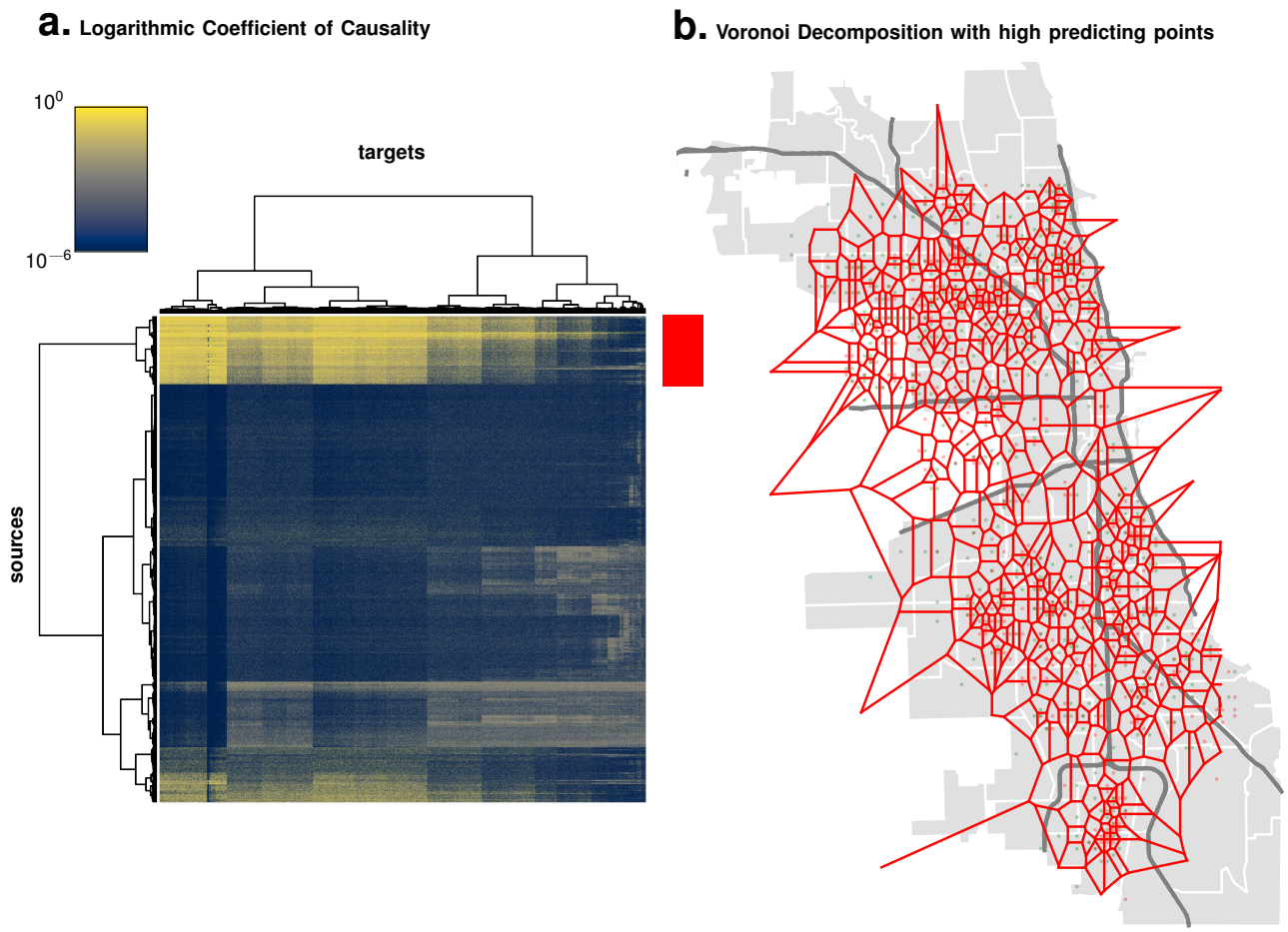

Fig. 10. **Automatic Neighborhood Decomposition Using Event Predictability** Computing a biclustering on the source-vs-target influence matrix (panel A) isolates a set of spatial tiles that are, on average, good predictors for all other tiles. Using this set, we use a Voronoi decomposition of the city (Panel B), which realizes an automatic spatial decomposition of the urban space, driven by event predictability.

Supplementary Text: Precise Event-level Prediction of Urban Crime Reveals Signature of Enforcement Bias

Victor Rotaru[1,3], Yi Huang[1], Timmy Li[1,3], James Evans[2,5,6] and Ishanu Chattopadhyay,[1,4,5]★

[1]Department of Medicine, University of Chicago, Chicago, IL 60637, USA
[2]Department of Sociology, University of Chicago, Chicago, IL 60637, USA
[3]Department of Computer Science, University of Chicago, Chicago, IL 60637, USA
[4]Committee on Quantitative Methods in Social, Behavioral, and Health Sciences, University of Chicago, Chicago, IL 60637, USA
[5]Committee on Genetics, Genomics & Systems Biology, University of Chicago, Chicago, IL 60637, USA
[6]Santa Fe Institute, Santa Fe NM 87501, USA

★To whom correspondence should be addressed: e-mail: ishanu@uchicago.edu.

CONTENTS

1 ALGORITHM PSEUDOCODE

Algorithm 1: Granger Net

Data:
- a set of sequence $\{x_i : i = 1, \ldots, N\}$ of length n;
- a hyperparameter $0 < \varepsilon < 1$;
- a model inference length $n_0 < n$;
- a maximal delay $\Delta_{\max}$;
- a threshold coefficient of causal dependence γ_0 for admissible models;

Result: A set of XPFSA models and a set of scalar weights for each target $r \in \{1, \ldots, N\}$.

/* **Infer models** */

Let $\mathcal{M}_r = \emptyset$ be the set of admissible models for each target $r \in \{1, \ldots, N\}$;
for *each delay* $\Delta = 1, \ldots, \Delta_{max}$ **do**
  for *each source* $s = 1, \ldots, N$ *and target* $r = 1, \ldots, N$ **do**
    Let $x_{\text{in}} = (x_s)_1^{n_0 - \Delta}$;
    Let $x_{\text{out}} = (x_r)_{\Delta+1}^{n_0}$;
    Calculate PFSA $G = \mathbf{GenESeSS}(x_{\text{in}}, \varepsilon)$;
    Calculate XPFSA $H^s_{r,\Delta} = \mathbf{xGenESeSS}(x_{\text{out}}, \varepsilon)$;
    Let $\gamma^s_{r,\Delta} = \mathbf{coefCausalDependence}(G, H^s_{r,\Delta})$;
    if $\gamma^s_{r,\Delta} \geq \gamma_0$ **then**
      Let $\mathcal{M}_r = \mathcal{M}_r \cup \left\{H^s_{r,\Delta}\right\}$;

/* **Learn scalar weights** */

for *each target* $r = 1, \ldots, N$ **do**
  Let $I_r = \left\{(s, \Delta) : \text{ there is a model } H^s_{r,\Delta} \in \mathcal{M}_r\right\}$;
  for *each timestamp* $t = 1, \ldots, n - n_0$ **do**
    Let $\mathbf{x}_t$ be a vector with index set I_r;
    for *each pair* $(s, \Delta) \in I_r$ **do**
      Let x_{in} the length l sub-sequence of x_s that ends in the $(n_0 + t - \Delta)$-th entry;
      Let the entry of $\mathbf{x}_t[s, \Delta] = \mathbf{predict}\left(H^s_{r,\Delta}, x_{\text{in}}\right)$;
      Let $y_t = x_r[n_0 + t]$;
  Let X the matrix with the t-th row being $\mathbf{x}_t$;
  Let $\mathbf{y}$ be the vector with the t-th entry being y_t;
  Initialize a suitable regressor **Reg**;
  Get scalar weights $\mathbf{w}_r = \left(w^s_{r,\Delta}\right)_{(s,\Delta)\in I_r} = \mathbf{Reg}(X, \mathbf{y})$;
return $\{(\mathcal{M}_r, \mathbf{w}_r) : r = 1, \ldots, N\}$;

Algorithm 2: **GenESeSS**

Data: A sequence x over alphabet Σ, $0 < \varepsilon < 1$
Result: State set Q, transition map δ, and transition probability $\widetilde{\pi}$

/* **Step One: Approximate ε-synchronizing sequence** */

Let $L = \left\lceil \log_{|\Sigma|} 1/\varepsilon \right\rceil$;
Calculate the **derivative heap** $\mathcal{D}^x_\varepsilon$ equaling $\left\{\hat{\phi}^x_y : y \text{ is a sub-sequence of } x \text{ with } |y| \leq L\right\}$;
Let $\mathcal{C}$ be the convex hull of D^x_ε;
Select x_0 with $\hat{\phi}^x_{x_0}$ being a vertex of $\mathcal{C}$ and has the highest frequency in x;

/* **Step Two: Identify transition structure** */

Initialize $Q = \{q_0\}$;
Associate to q_0 the **sequence identifier** $x^{\text{id}}_{q_0} = x_0$ and the probability vector $d_{q_0} = \hat{\phi}^x_{x_0}$;
Let $\widetilde{Q}$ be the set of states that are just added and initialize it to be Q;
while $\widetilde{Q} \neq \emptyset$ **do**
  Let $Q_{\text{new}} = \emptyset$ be the set of new states;
  for $(q, \sigma) \in \widetilde{Q} \times \Sigma$ **do**
    Let $x = x^{\text{id}}_q$ and $d = \hat{\phi}^x_{x\sigma}$;
    if $\|d - d_{q'}\|_\infty < \varepsilon$ *for some* $q' \in Q$ **then**
      Let $\delta(q, \sigma) = q'$;
    else
      Let $Q_{\text{new}} = Q_{\text{new}} \cup \{q_{\text{new}}\}$ and $Q = Q \cup \{q_{\text{new}}\}$;
      Associate to q_{new} the sequence identifier $x^{\text{id}}_{q_{\text{new}}} = x\sigma$ and the probability vector $d_{q_{\text{new}}} = d$;
      Let $\delta(q, \sigma) = q_{\text{new}}$;
  Let $\widetilde{Q} = Q_{\text{new}}$;
Take a strongly connected subgraph of the labeled directed graph defined by Q and δ, and denote the vertex set of the subgraph again by Q;

/* **Step Three: Identify transition probability** */

Initialize counter $N[q, \sigma]$ for each pair $(q, \sigma) \in Q \times \Sigma$;
Choose a random starting state $q \in Q$;
for $\sigma \in x$ **do**
  Let $N[q, \sigma] = N[q, \sigma] + 1$;
  Let $q = \delta(q, \sigma)$;
Let $\widetilde{\pi}(q) = [\![(N[q, \sigma])_{\sigma\in\Sigma}]\!]$;
return $Q, \delta, \widetilde{\pi}$;

Algorithm 3: `xGenESeSS`

Data: A sequence x_{in} over alphabet Σ_{in}, a sequence x_{out} over alphabet Σ_{out}, and $0 < \varepsilon < 1$
Result: State set R, transition map η, and output probability χ

/* **Step One: Approximate ε-synchronizing sequence** */
Let $L = \left\lceil \log_{|\Sigma_{\text{in}}|} 1/\varepsilon \right\rceil$;
Calculate **cross derivative heap** $\mathcal{D}_{\varepsilon}^{x_{\text{in}}, x_{\text{out}}}$ equaling $\left\{ \hat{\phi}_{y}^{x_{\text{in}}, x_{\text{out}}} \ : \ y \text{ is a sub-sequence of } x_{\text{in}} \text{ with } |y| \leq L \right\}$;
Let $\mathcal{C}$ be the convex hull $\mathcal{D}_{\varepsilon}^{x_{\text{in}}, x_{\text{out}}}$;
Select x_0 with $\hat{\phi}_{x_0}^{x_{\text{in}}, x_{\text{out}}}$ being a vertex of $\mathcal{C}$ and has the highest frequency in x;
/* **Step Two: Identify transition structure** */
Initialize $R = \{r_0\}$;
Associate to r_0 the **sequence identifier** $x_{r_0}^{\text{id}} = x_0$ and the probability vector $\chi(r_0) = \hat{\phi}_{x_0}^{x_{\text{in}}, x_{\text{out}}}$;
Let $\widetilde{R}$ be the set of states that are just added and initialize it to be R;
while $\widetilde{R} \neq \emptyset$ **do**
    Let $R_{\text{new}} = \emptyset$ be the set of new states;
    for $(r, \sigma) \in \widetilde{R} \times \Sigma_{in}$ **do**
        Let $x = x_r^{\text{id}}$ and $d = \hat{\phi}_{x\sigma}^{x_{\text{in}}, x_{\text{out}}}$;
        if $\|d - \chi(r')\|_{\infty} < \varepsilon$ *for some* $r' \in R$ **then**
            Let $\eta(r, \sigma) = r'$;
        else
            Let $R_{\text{new}} = R_{\text{new}} \cup \{r_{\text{new}}\}$ and $R = R \cup \{r_{\text{new}}\}$;
            Associate to r_{new} the sequence identifier $x_{r_{\text{new}}}^{\text{id}} = x\sigma$ and the probability vector $\chi(r_{\text{new}}) = d$;
            Let $\eta(r, \sigma) = r_{\text{new}}$;
    Let $\widetilde{R} = R_{\text{new}}$;
Take a strongly connected subgraph of the labeled directed graph defined by R and η, and denote the vertex set of the subgraph again by R;
/* **Step Three: Identify output probability** */
Initialize counter $N[r, \tau]$ for each pair $(r, \tau) \in R \times \Sigma_{\text{out}}$;
Choose a random starting state $r \in R$;
for $i \in 1, \ldots, |x_{in}|$ **do**
    Let σ_i be the i-th symbol in x_{in} and τ_i be the i-th symbol in x_{out};
    Let $N[r, \tau_i] = N[r, \tau_i] + 1$;
    Let $r = \eta(r, \sigma_i)$;
Let $\chi(r) = \left[\!\left[(N[r, \tau])_{\tau \in \Sigma_{\text{out}}} \right]\!\right]$;
return R, η, χ;

2 THEORY OF PROBABILISTIC AUTOMATA

Granger Net is assembled from local models which are, in general, crossed probabilistic automata (XPFSA).

The construction of a Granger Net consists of two steps: 1) local model generation and network pruning and 2) local model aggregation for comprehensive prediction. Event prediction is accomplished by aggregating these local activations via a local regressor. No global optimization of these aggregation function is acrried out.

The model generation step of Granger Net is accomplished by the algorithms **GenESeSS** (See Algorithm 2) and **xGenESeSS** (See Algorithm 3). **xGenESeSS** produces XPFSA models that captures how the history of a source process influences the future of a target process. The Granger Net construction is described in Algorithm 1, and takes as input a set $\{x_s : s \in S\}$ of length-n time series, hyperparameters ε and $n_0 < n$ for local model inference, $\Delta_{\max}$ for maximum time delay, and γ_0 for thresholding admissible models. For each target sequence x_r, Granger Net outputs a set of admissible models $\mathcal{M}_r$ with a scalar weight for each model in $\mathcal{M}_r$ via model inference and pruning (line 1-10) and training of the aggregation weights (line 11-22).

Step 1: Model inference and pruning

The Granger Net framework models the influence from a source time series x_s on a target time series x_r at a particular time delay Δ by an XPFSA $H^s_{r,\Delta}$ (line 7). Thus, we infer $|S|\Delta_{\max}$ XPFSA models for each x_r which yields $|S|^2\Delta_{\max}$ models in total. Since the number of XPFSA models increases quadratically with the number of time series and strength of the links may vary, pruning low-performing models early is important for parsimony. Granger Net rejects models by thresholding on the *coefficient of causal dependence* $\gamma^s_{r,\Delta}$ of model $H^s_{r,\Delta}$ (line 8), which measures the strength of dependence of the output sequence on the input one. More specifically, we have

$$\gamma^s_{r,\Delta} = 1 - \frac{\text{uncertainty of the next output in } x_r \text{ with observation of } x_s}{\text{uncertainty of the next output in } x_r} \tag{1}$$

γ can be evaluated from the *synchronous composition* of the PFSA that models the input process (line 6) and the XPFSA that models the causal influence. Granger Net retains the model $H^s_{r,\Delta}$ if and only if $\gamma^s_{r,\Delta}$ is greater than a pre-specified threshold γ_0. At the conclusion of Step 1, Granger Net returns an admissible set of models

$$\mathcal{M}_r = \left\{H^s_{r,\Delta} : \gamma^s_{r,\Delta} > \gamma_0\right\} \tag{2}$$

for each $r \in S$.

Step 2: Train linear weights

In this step, we integrate the local models in x_r's admissible set for forecasting events in x_r. To do this, Granger Net trains a linear coefficient $\omega^s_{r,\Delta}$ for each $H^s_{r,\Delta} \in \mathcal{M}_r$ (line 22) so that the final prediction for x_r at time step h is equal to

$$\sum_{H^s_{r,\Delta} \in \mathcal{M}_r} \omega^s_{t,\Delta} H^s_{r,\Delta}\left((x_s)^{h-\Delta}\right), \tag{3}$$

where $(x_s)^{h-\Delta}$ is the truncation of x_s at $h-\Delta$. To compute the coefficients, we solve a regression problem $\texttt{Reg}(X, \mathbf{y})$ (line 22) for each $r \in S$ with the predictor variables being predictions $\mathbf{x}_t[s,\Delta]$ obtained by running each sequence $(x_s)^{n_0+t-\Delta}$ through $H^s_{r,\Delta}$ (line 17), and the outcome variable being $x_r[n_0+t]$, value of x_r at time n_0+t (line 18). Hence, the X matrix is the $(n-n_0) \times |\mathcal{M}_r|$ matrix with the entry indexed by $t, (s,\Delta)$ given by $\mathbf{x}_t[s,\Delta]$ and $\mathbf{y}$, the $(n-n_0)$-dimensional vector with the entry indexed by t given by $x_r[n_0+t]$. We can solve for the linear weights with any standard regressor.

Inference Algorithms

On line 6 and 7 of Algorithm 1, Granger Net calls subroutine **xGenESeSS**, which infers XPFSA as models of cross-dependencies between processes. Here, we establish the correctness of **GenESeSS**.

The inference algorithm for PFSA is called **GenESeSS** for Generator Extraction Using Self-similar Semantics. The PFSA model is based on the concept of the *causal state*. A dynamical system reaches the same causal state via distinct paths if the futures are statistically indistinguishable. More precisely, each process over an alphabet Σ of size m gives rise naturally to an m-ary tree with the nodes at level d being sequences of length d, and the edge from the node x to $x\sigma$, $\sigma \in \Sigma$, labeled by $Pr(\sigma|x)$ – the probability of observing σ as the next output after x. By the definition of causal state, if two subtrees are identical with respect to edge labels, then their roots are sequences that lead the system to the *same* causal state. Identifying all the roots of identical subtrees induces a *finite* automaton structure whose unique strongly connected component is the generating model of the process.

Definition 1 (Probabilistic Finite-State Automaton (PFSA))**.** *A PFSA G is a quadruple $(Q, \Sigma, \delta, \widetilde{\pi})$, where Q is a finite set, Σ is a finite alphabet, $\delta : Q \times \Sigma \to \Sigma$ is called the transition map, and $\widetilde{\pi} : Q \to \mathbf{P}_\Sigma$, where $\mathbf{P}_\Sigma$ is the space of probability distributions over Σ, is called the transition probability.*

Step 2 of Algorithm 2 (line 5-19) is an implementation this subtree "stitching" approach under finiteness of input data.

Note that the criterion for "stitching" two subtrees with roots x and x' is that their edge labels are identical for *all* depths, which translates to $p(y|x) = p(y|x')$ for sequence y of all lengths. The criterion is not verifiable with finite data, and hence **GenESeSS** identifies two subtrees if they agree on depth one. Defining *symbolic derivative* ϕ_x to be the vector with the entry indexed by σ given by $p(\sigma|x)$, **GenESeSS** identifies x and x' if $\phi_x = \phi_{x'}$. This approach works well under the assumption that the target PFSA is in *general position*, meaning that different causal states have distinct symbolic derivatives. In practice, **GenESeSS** uses *empirical symbolic derivative* defined below to approximate ϕ_x. Let x be an input sequence of finite length, the empirical symbolic derivative $\hat{\phi}_y^x$ of a sub-sequence y of x is a probability vector with the entry indexed by σ given by

$$\hat{\phi}_y^x(\sigma) = \frac{\text{number of } y\sigma \text{ in } x}{\text{number of } y \text{ in } x} \tag{4}$$

GenESeSS identifies two sequences (line 12) if their *empirical* symbolic derivatives are within an ε-neighborhood of each other for certain $\varepsilon > 0$.

For simplicity, we first illustrate how **GenESeSS** solves the transition structure of the target PFSA from a sample path x generated from a process of Markov order k. Assuming the x_0 produced by Step 1 (line 4) is λ, the empty sequence, **GenESeSS** starts by calculating $\hat{\phi}_\lambda^x$, *i.e.*, the empirical distribution on Σ, and records λ as the identifier of the first state. Then, **GenESeSS** appends λ with each $\sigma \in \Sigma$, and calculates $\hat{\phi}_\sigma^x$. By the general position assumption and assuming x is long enough, with high probability, no $\hat{\phi}_\sigma^x$ is within an ε-neighborhood of $\hat{\phi}_{\sigma'}^x$ for $\sigma \neq \sigma'$, and hence each σ is recorded as the identifier for a new state. In fact, **GenESeSS** will keep on appending symbols to identifiers of stored states and adding new states until it reaches a sequence of length $k+1$. Assuming $y = \sigma_1 \cdots \sigma_k \sigma_{k+1}$, since the process is of order k, we have $\phi_y = \phi_z$ for $z = \sigma_2 \cdots \sigma_{k+1}$, and hence, with high probability, $\hat{\phi}_y^x$ and $\hat{\phi}_z^x$ can be within an ε-neighborhood of each other given long enough input x. In this case, **GenESeSS** identifies the state represented by y with that of z. In fact, **GenESeSS** will identify all states represented by sequences of length $k+1$ to some previously-stored states. And since no new states can be found, **GenESeSS** exits the loop on line 8 after iteration $k+1$. Taking the strongly connected component on line 19, **GenESeSS** gets the correct transition structure.

However, not all processes generated by PFSA have finite Markov order. For such cases, Step 2 of **GenESeSS** will never exit in theory, since there exists no $n \in \mathbb{N}$ such that every causal state is visited for sequences with length $\leq n$. And if we implement an artificial exit criterion, the model inferred might be unnecessarily large, and have hard-to-model approximations. We address this issue via the notion of synchronization – the ability to identify that we are localized or synchronized to a particular state despite being uncertain of the initial state.

In Step 1 of Algorithm 2 (line 1-4), **GenESeSS** finds an almost synchronizing sequence, which allows **GenESeSS** to distill a structure that is similar to that of the finite Markov order cases, and thus carry out the subtree "stitching" procedure described before. A sequence x is *synchronizing* if *all* sequences that end with the suffix x terminates on the same causal state. A process is synchronizable if it has a synchronizing sequence, and a PFSA is *synchronizable* if the process it generates is synchronizable. The structure of the "graph" of a perfectly synchronizable PFSA is that of a co-final automata[1].

A sequence x is *ε-synchronizing*[2] to the state q if the distribution $\wp_x$ on the state set Q induced by x satisfies $\|\wp_x - \mathbf{e}_q\|_\infty < \varepsilon$, where $\mathbf{e}_q$ is the base vector with 1 on the entry indexed by q and 0 elsewhere. The importance of ε-synchronizing sequence is twofold: 1) since $\phi_x^T = \wp_x^T \widetilde{\Pi}$, where $\widetilde{\Pi}$ is the $|Q| \times |\Sigma|$ matrix with the row indexed by q given by $\widetilde{\pi}(q)$, a $\wp_x$ close to $\mathbf{e}_q$ give rise to a ϕ_x close to $\widetilde{\pi}(q)$. And 2) although sequences prefixed by an ε-synchronizing sequence to a state q may not remain ε-synchronizing to state q, they are close to q *on average*.

To find an almost synchronizing sequence algorithmically[2], **GenESeSS** first calculates the convex hull of symbolic derivatives of subsequences of x up to length L (line 1-3), and then selects a sequence x_0 whose symbolic derivative is a vertex of the convex hull (line 4). Since the convex hull of $\left\{\phi_x : x \in \Sigma^L\right\}$ is a linear projection of the convex hull $\left\{\wp_G(x) : x \in \Sigma^L\right\}$ via $\widetilde{\Pi}$, we can expect sequence x with ϕ_x being a vertex of the convex hull of $\left\{\phi_x : x \in \Sigma^L\right\}$ to be a good candidate for an almost synchronizing sequence.

The corresponding inference algorithm for XPFSA is called **xGenESeSS**, which takes as input two sequences x_{in}, x_{out}, and a hyperparameter ε, and outputs an XPFSA in a manner very similar to the inference algorithm of PFSA.

While a PFSA models how the past of a time series influences its own future, a XPFSA models how the past of an input time series influences the future of an output time series. Hence, while in the SSC algorithm of PFSA, we identify sequences if they lead to futures that are statistically indistinguishable, in the SSC algorithm of XPFSA, we identify sequences if they lead to the same future distribution of the *output*.

Definition 2 (Crossed Probabilistic Finite-State Automaton (XPFSA))**.** *A crossed probabilistic finite-state automaton is specified by a quintuple $(\Sigma_{in}, R, \eta, \Sigma_{out}, \chi)$, where Σ_{in} is a finite input alphabet, R is a finite state set, η is a partial function from $R \times \Sigma_{in}$ to R called transition map, Σ_{out} is a finite output alphabet, and χ is a function from R to $\mathbf{P}_{\Sigma_{out}}$ called output probability map, where $\mathbf{P}_{\Sigma_{out}}$ is the space of probability distributions over Σ_{out}. In particular, $\chi(r, \tau)$ is the probability of generating $\tau \in \Sigma_{out}$ from a state $r \in R$.*

Note that a XPFSA has no transition probabilities defined between states as a PFSA does. The XPFSA in the example

has a binary input alphabet and an output alphabet of size 3. The bar charts next to the 4 states of the XPFSA indicate the output probability distributions. To generate a sample path, an XPFSA requires an input sequence over its input alphabet.

Similar to the PFSA construction approach, here we compute the *cross symbolic derivative*, which is the ordered tuple $Pr(\tau|x)$, with $\tau \in \Sigma_{\text{out}}$ and a sequence x over Σ_{in}. We compute the empirical approximation of the cross symbolic derivative from sequences x_{in} and x_{out} as:

$$\hat{\phi}_y^{x_{\text{in}},x_{\text{out}}}(\tau) = \frac{\text{number of } \tau \text{ in } x_{\text{out}} \text{ after } y \text{ transpires in } x_{\text{in}}}{\text{number of sub-sequence } y \text{ in } x_{\text{in}}} \tag{5}$$

Thus, **xGenESeSS** is almost identical to **GenESeSS** except that, in Step 1, **xGenESeSS** finds an almost synchronizing sequence based on cross symbolic derivatives, and in Step 2, identifies the transition structure based on the similarity between cross symbolic derivatives. Arguments for establishing the effectiveness of **GenESeSS** carry over to **xGenESeSS** with empirical symbolic derivative replaced by empirical cross symbolic derivative.

3 Software Availability & Repository

Software for the cynet implementation, with instructions for installation and quick-start examples, is available at https://pypi.org/project/cynet/

References

[1] Ito, M. & Duske, J. On cofinal and definite automata. *Acta Cybernetica* **6**, 181–189 (1983).
[2] Chattopadhyay, I. & Lipson, H. Abductive learning of quantized stochastic processes with probabilistic finite automata. *Philos Trans A* **371**, 20110543 (2013).

Figures

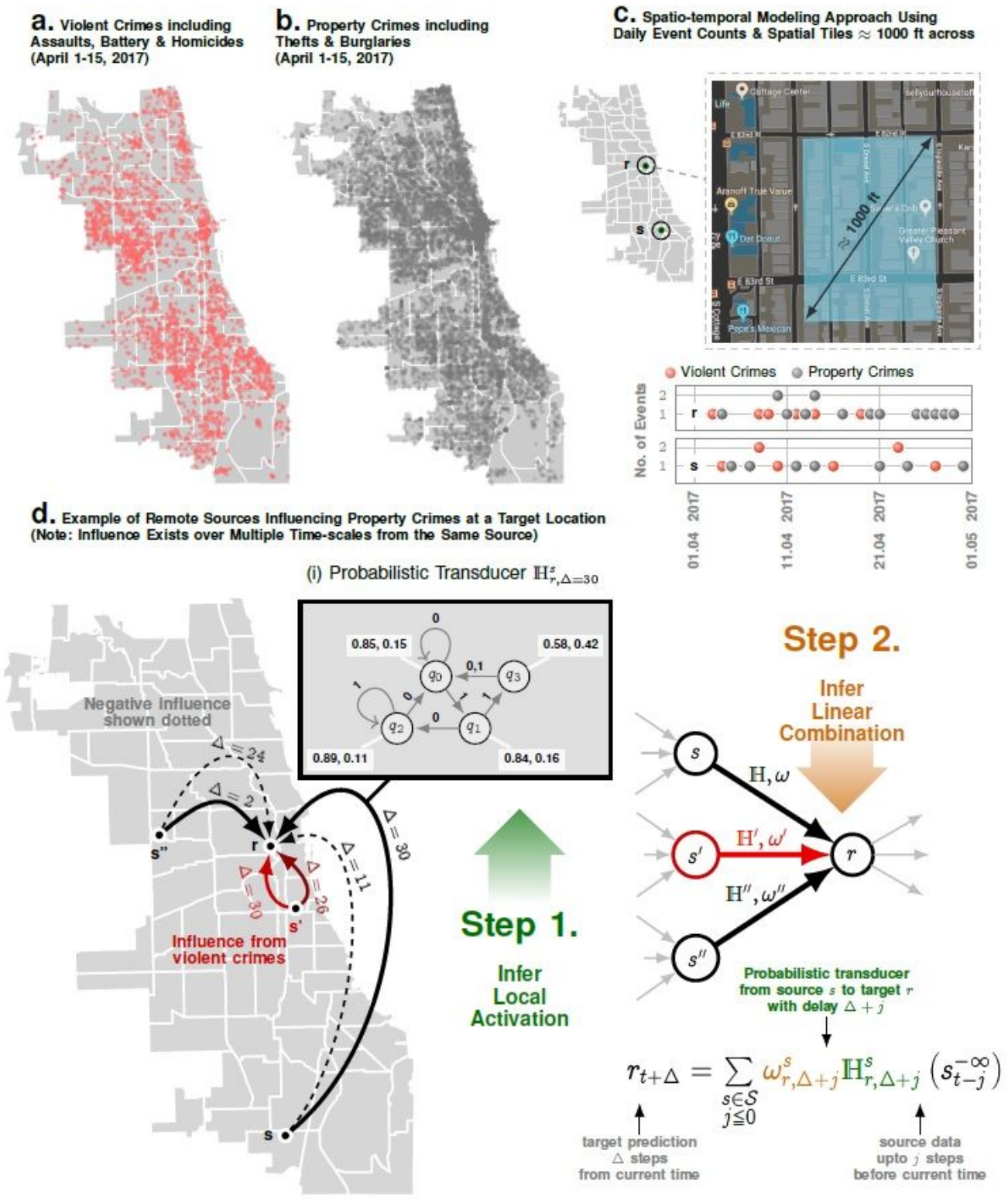

Figure 1

Crime Data & Modeling Approach. a and b show the recorded infractions within the 2 week period between April 1 and 15 in 2017. Plate c illustrates our modeling approach: We break city into small spatial tiles approximately 1.5 times the size of an average city block, and compute models that capture

multi-scale dependencies between the sequential event streams recorded at distinct tiles. In this paper, we treat violent and property crimes separately, and show that these categories have intriguing cross-dependencies. Plate d illustrates our modeling approach. For example, to predict property crimes at some spatial tile r, we proceed as follows: Step 1) we infer the probabilistic transducers that estimate event sequence at r by using as input the sequences of recorded infractions (of different categories) at potentially all remote locations (s; s 0 ; s 00 shown), where this predictive influence might transpire over different time delays (a few shown on the edges between s and r). Step 2) Combine these weak estimators linearly to minimize zero-one loss. The inferred transducers can be thought of as inferred local activation rules, which are then linearly composed, reversing the approach of linearly combining input and then passing through fixed activation functions in standard neural net architectures. The connected network of nodes (variables) with probabilistic transducers on the edges comprises the Granger Network.

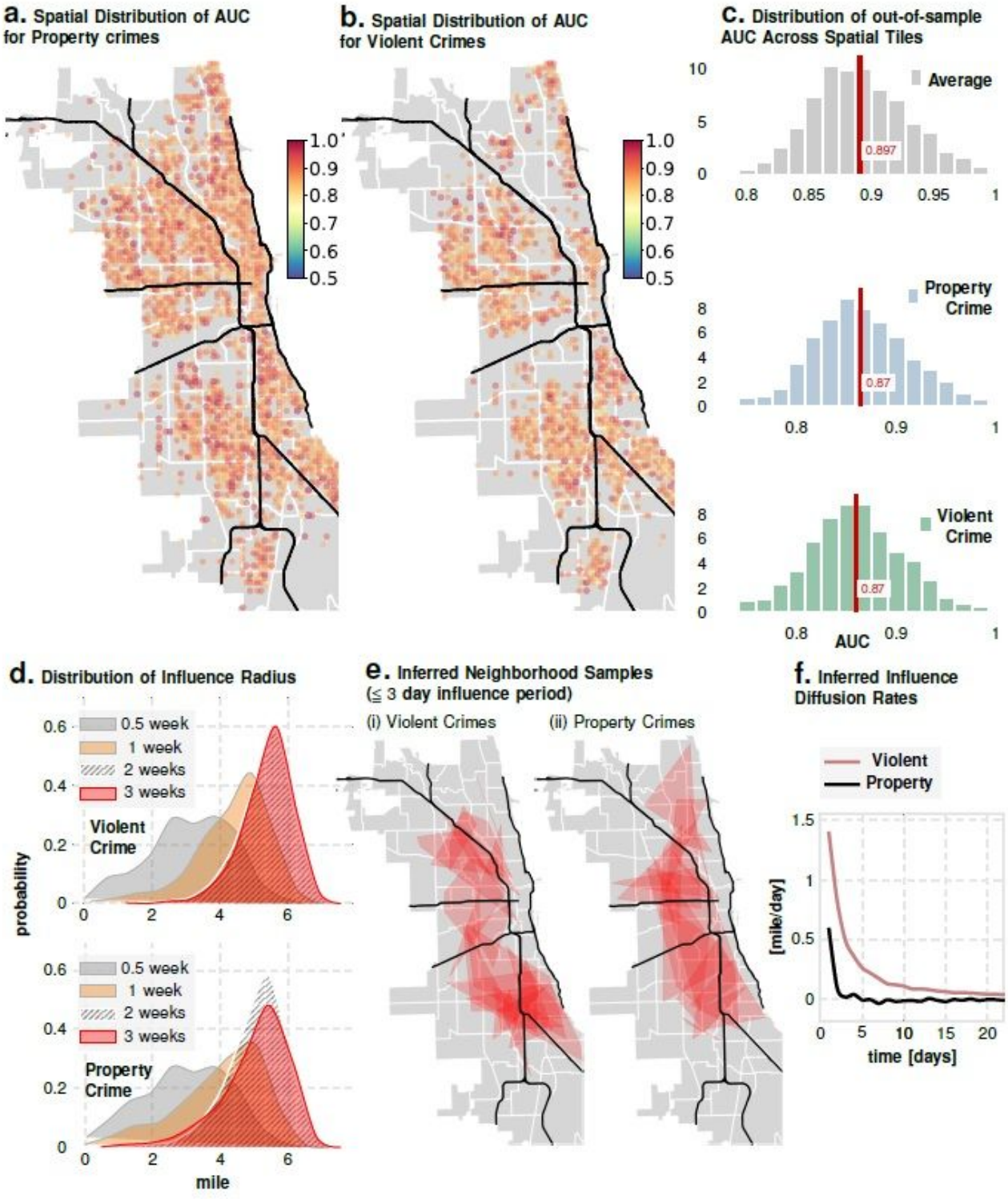

Figure 2

Predictive Performance of Granger Nets. a an b illustrate the out-of-sample area under the receiver operating characteristics curve (AUC) for predicting violent and property crimes respectively. The prediction is made a week in advance, and the event is registered as a successful prediction if we get a hit within +1 day of the predicted date. c illustrates the distribution of AUC on average, individually for violent and property crimes. Our mean AUC is close to 90%. Panels d-f shows influence Diffusion & Perturbation

Space. If we are able to infer a model that is predicts event dynamics at a specific spatial tile (the target) using observations from a source tile + days in future, then we say the source tile is within the influencing neighborhood for the target location with a delay of D. d illustrates the spatial radius of influence for 0.5, 1, 2 and 3 weeks, for violent (upper panel) and property crimes (lower panel). Note that the influencing neighborhoods, as defined by our model, are large and approach a radius of 6 miles. Given the geometry of the City of Chicago, this maps to a substantial percentage of the total area of urban space under consideration, demonstrating that crime manifests demonstrable long-range and almost city-wide influence. e illustrates the extent of a few inferred neighborhoods at time delay of at most 3 days. f illustrates the average rate of influence diffusion measured by number of predictive models inferred that transduce influence as we consider longer and longer time delays. Note that the rate of influence diffusion falls rapidly for property crimes, dropping to zero in about a week, whereas for violent crimes, the influence continues to diffuse even after three weeks.

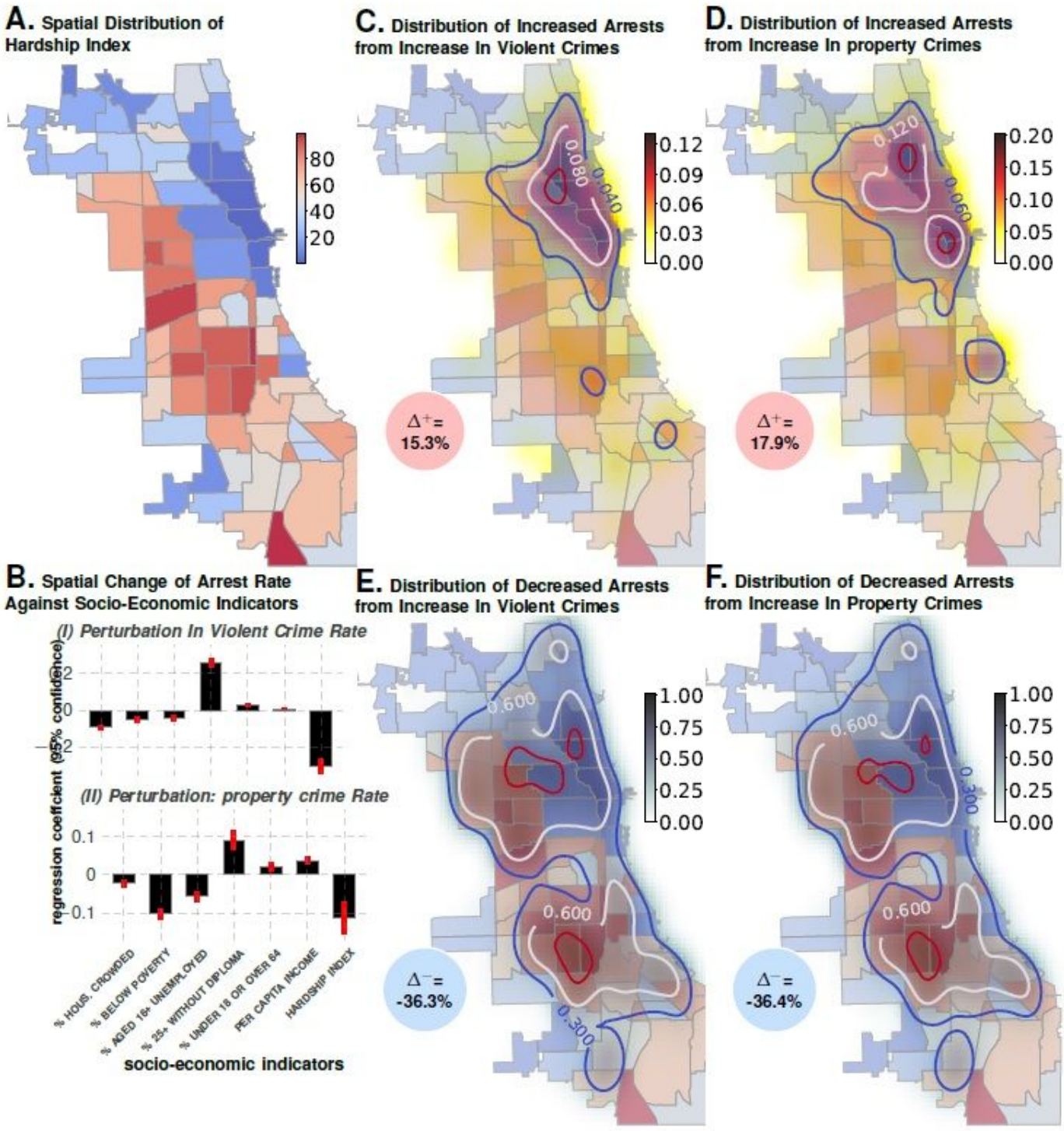

Figure 3

Estimating Bias. a illustrates the distribution of hardship index (see SI). c, d, e, and f suggest biased response to perturbations in crime rates. With a 10% increase in violent or property crime rates, we see an approximately a 30% decrease in arrests when averaged over the city. The spatial distribution of locations that experience a positive vs. negative change in arrest rate reveals a strong preference favoring wealthy locations. If neighborhoods are doing better socio-economically, increased crime predicts increased arrests. A strong converse trend is observed in predictions for poor and disadvantaged neighborhoods,

suggesting that under stress, wealthier neighborhoods drain resources from their disadvantaged counterparts. b illustrates this more directly via a multi-variable regression, where hardship index is seen to make a strong negative contribution.

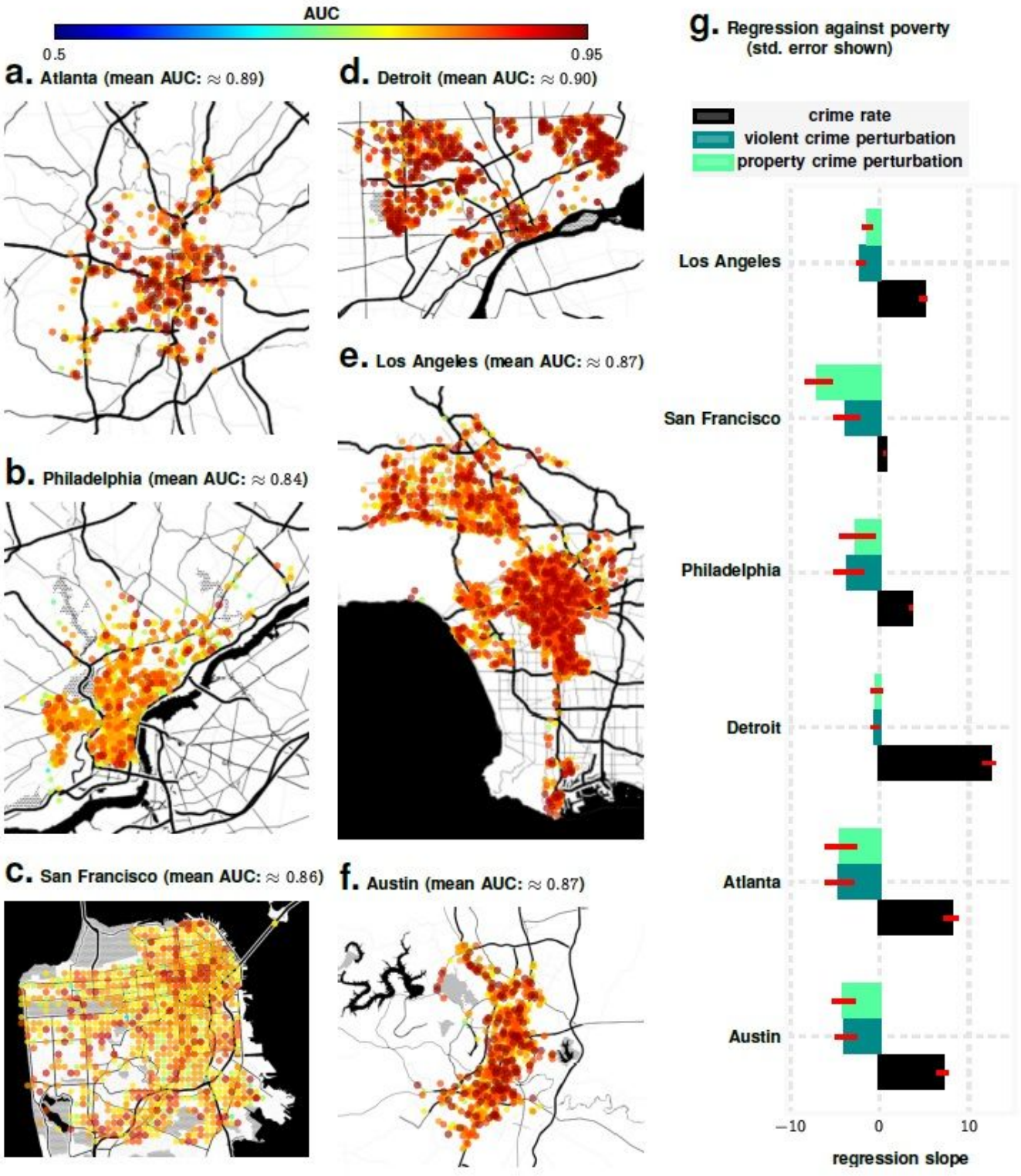

Figure 4

Prediction of property and violent crimes across major US cities and dependence of perturbation response on socio-economic status of local neighborhoods. Panels a-f illustrate the AUCs achieved in six

major US cities. These cities were chosen on the basis of the availability of detailed event logs in the public domain. All of these cities show comparably high predictive performance. Panel g illustrates the results obtained by regressing crime rate and perturbation response against SES variables (shown here for poverty, as estimated by the 2018 US census). We note that while crime rate typically goes up with increasing poverty, the number of events observed one week after a positive perturbation of 5-10% increase in crime rate is predicted to fall with increasing poverty. We suggest that this decrease is explainable by reallocation of enforcement resources disproportionately, away from disadvantaged neighborhoods in response to increased event rates, which leads to smaller number of reported crimes.

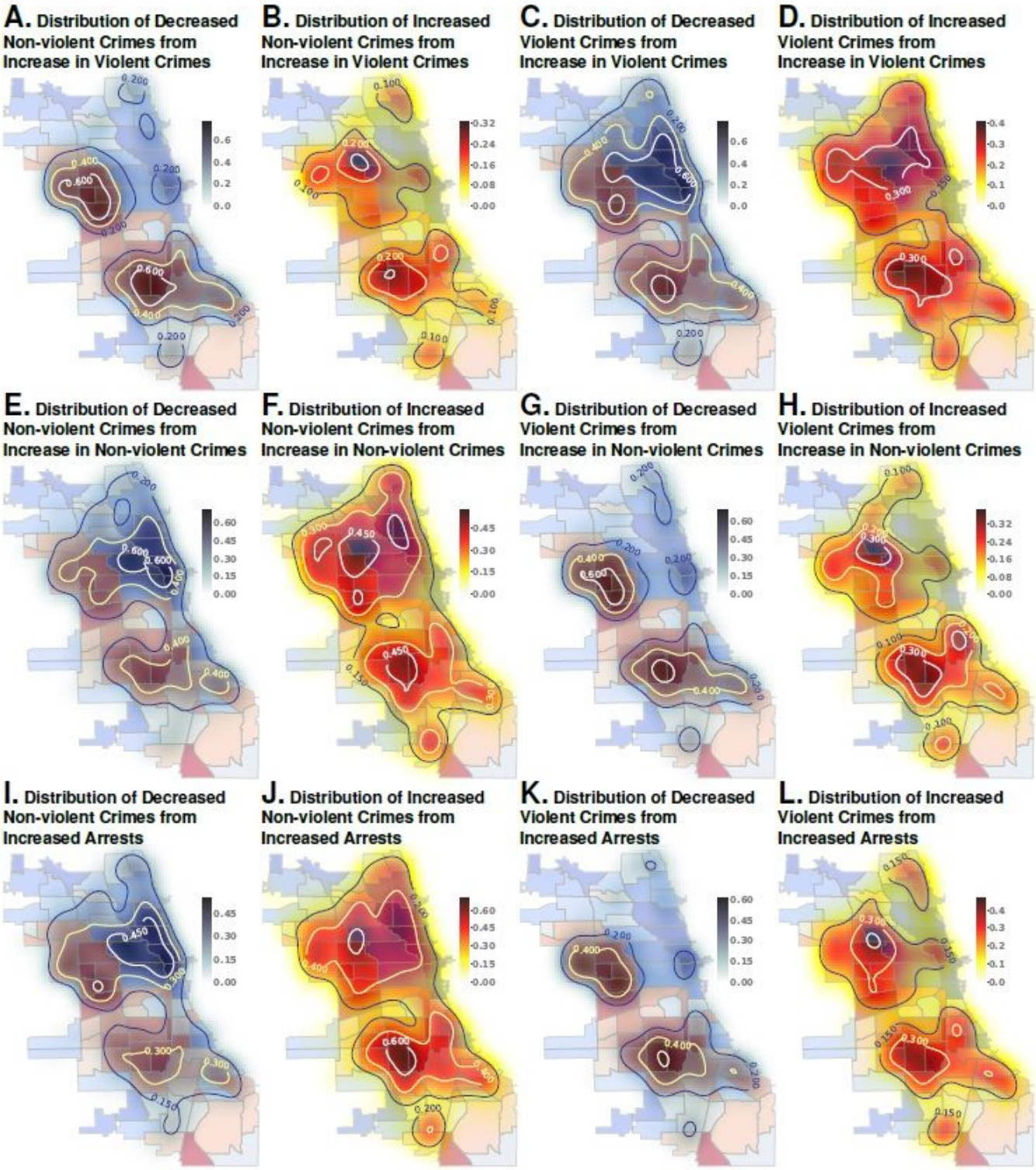

Figure 5

Perturbation Effects Across Variables. We see that the decrease of violent crimes from increase of property crimes are localized in disadvantaged neighborhoods (panel g). Similarly, the decrease of property crimes from increase of violent crimes is also localized to disadvantaged neighborhoods (panel a), as well as the decreased violent crimes from increased arrests (panel k). We see a weaker localization for the corresponding increases in crime rates under similar perturbations. Looking at other pairs of

variables under perturbation (rest of the panels), we generally do not see a very prominent correspondence with the distribution of socio-economic indicators. It seems crimes (and particulalrly violent crimes) are easier to dampen in lcales with high existing crime rates, which is desirable result. But such conclusions are currently confounded by SES variables, and futher work is needed to investigate these effects more thoroughly.

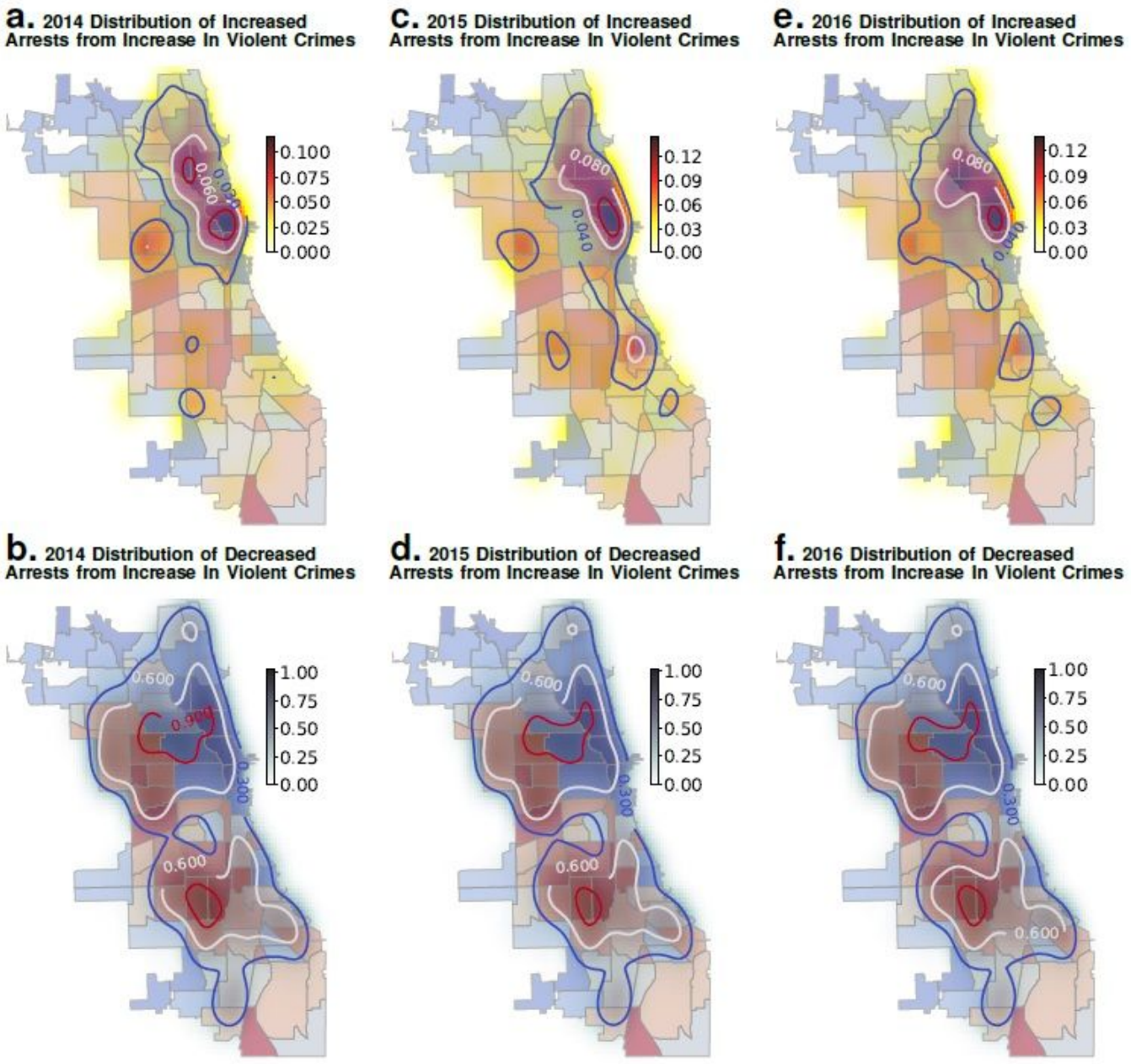

Figure 6

Stability of Suburban Bias over Years (Violent Crimes). We show that the nature of the perturbation response shown in Fig. 3 in the main text holds true for earlier years as well: panels a and b correspond to year 2014, c and d correspond to 2015 and e and f correspond to year 2016, all of which follow the same pattern shown in Fig. 3 in the main text.

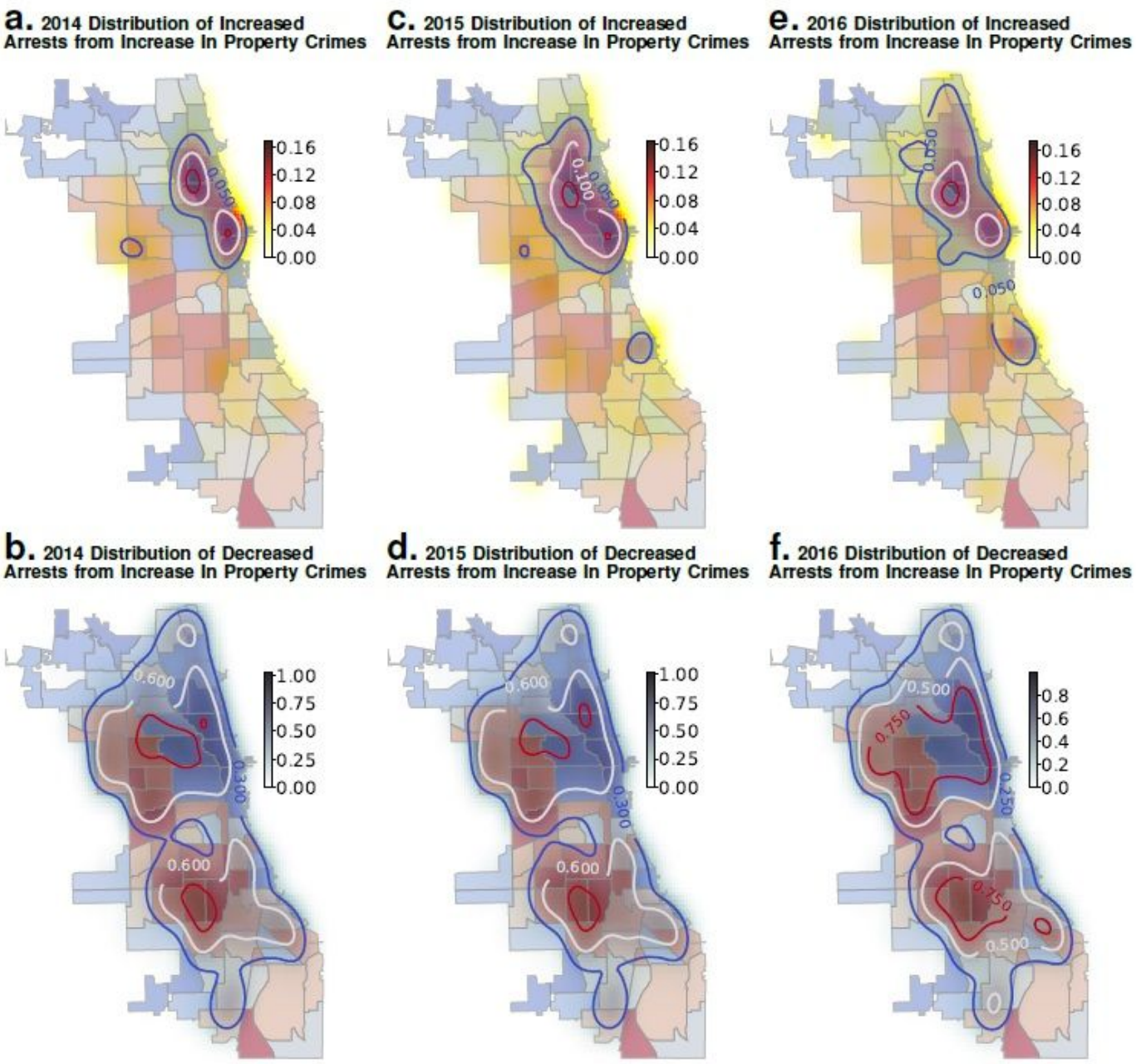

Figure 7

Stability of Suburban Bias over Years (Property Crimes). We show that the nature of the perturbation response shown in Fig. 3 in the main text holds true for earlier years as well: panels a and b correspond to year 2014, c and d correspond to 2015 and e and f correspond to year 2016, all of which follow the same pattern shown in Fig. 3 in the main text.

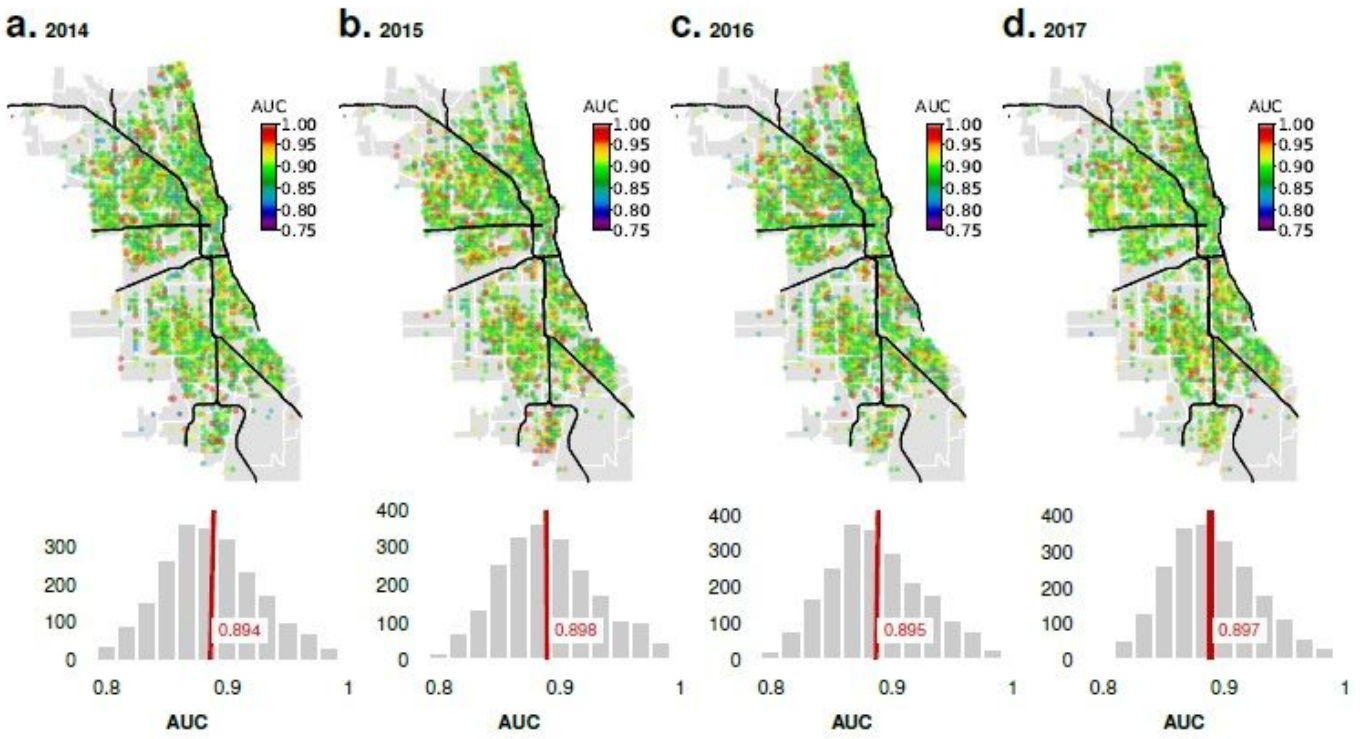

Figure 8

Out of Sample Predictive Performance over the Years. We show that the predictive performance is very stable, and variation in mean AUC is limited to the third place of decimal, at least when analyzing the last few years (4 years shown).

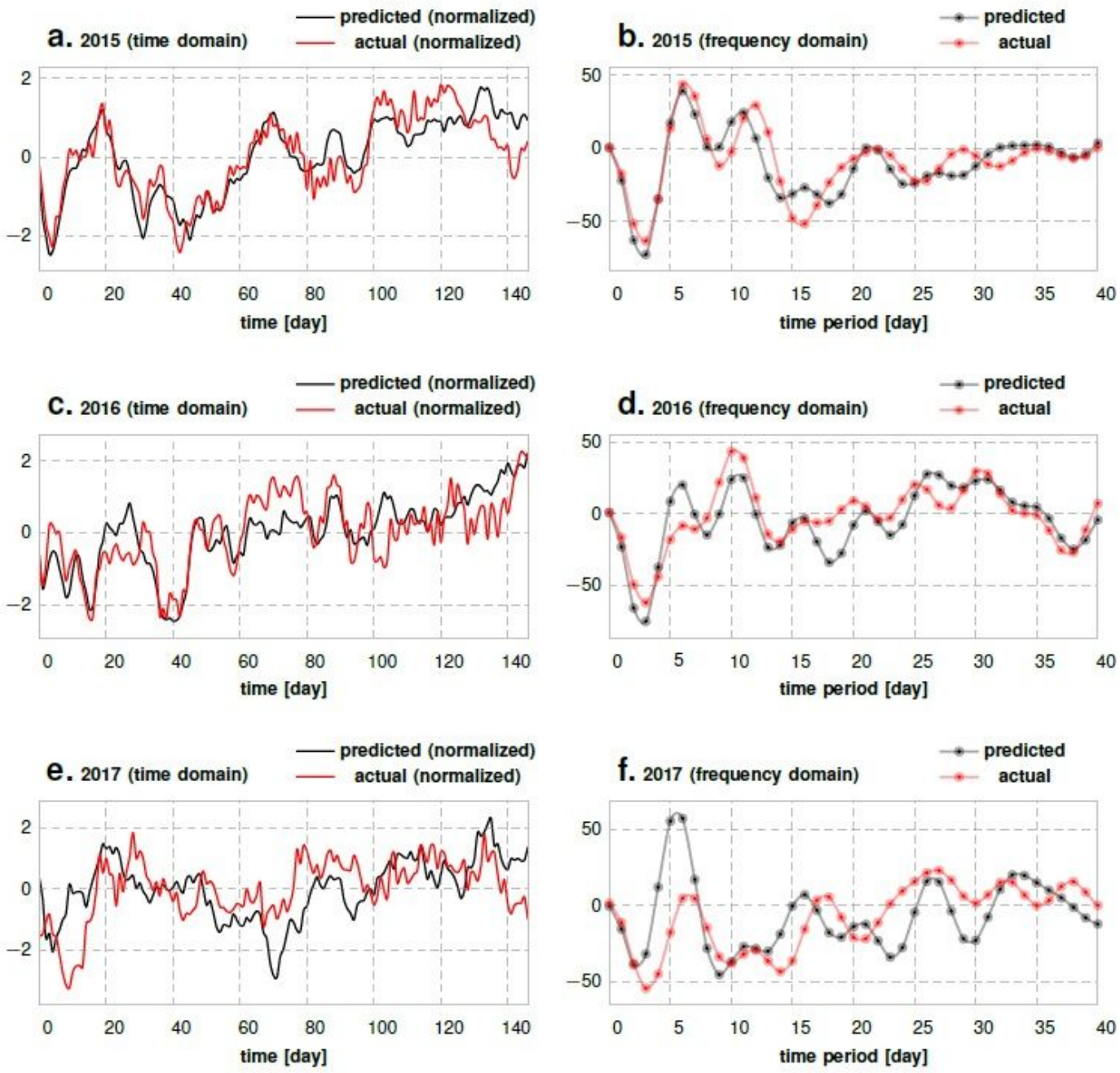

Figure 9

Comparison of Predicted vs Actual Sample Paths in Time and Frequency Domains. Panels a, c and e show that the predicted and actual sample paths are pretty close for different years, when compared over the first 150 days of each year. Panels b, d and f show that the Fourier coefficients match up pretty well as well. More importantly, while our models do not explicitly incorporate any periodic elements that are being tuned, we still manage to capture the weekly, (approximately) biweekly and longer periodic regularities.

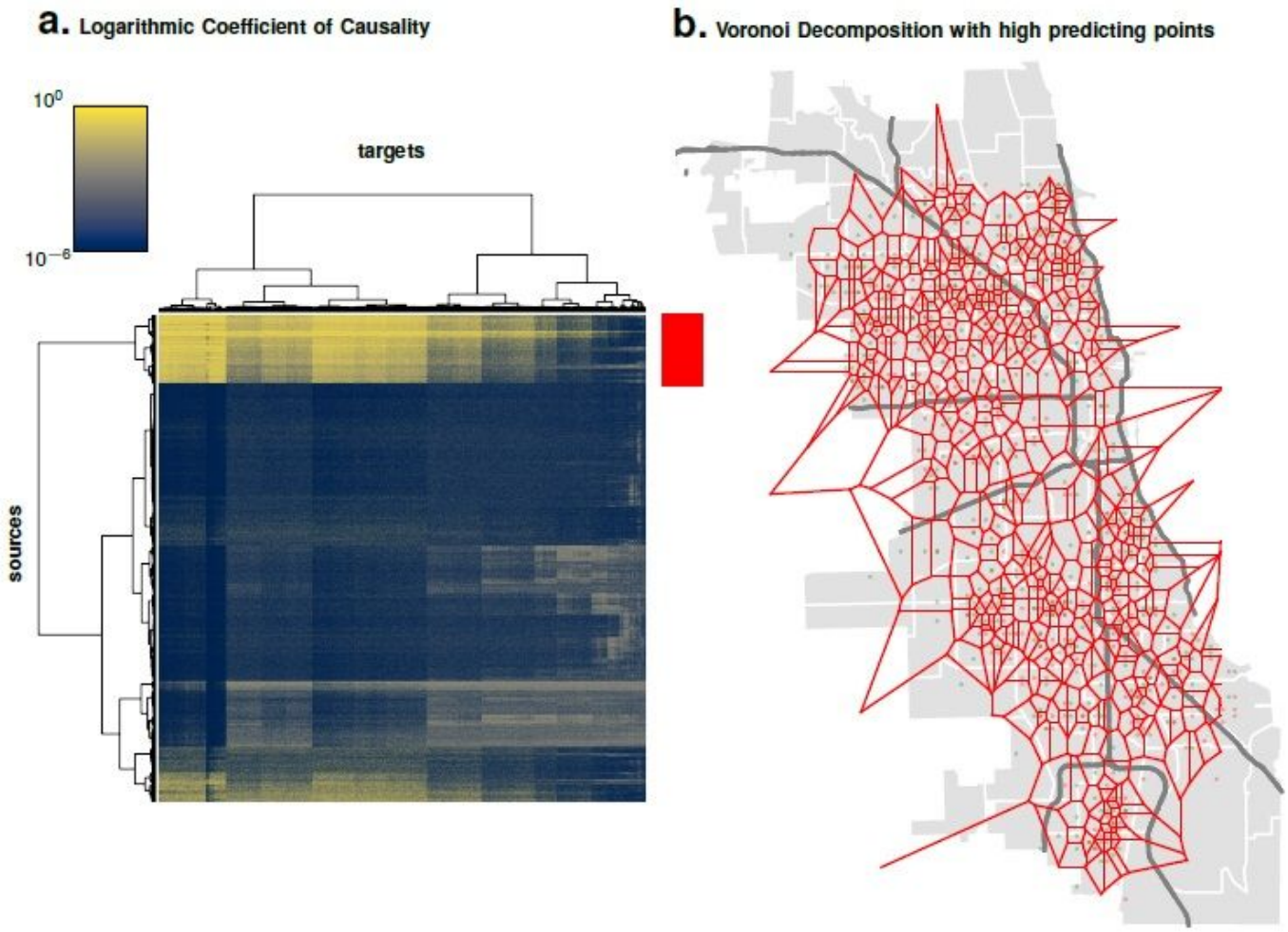

Figure 10

Automatic Neighborhood Decomposition Using Event Predictability Computing a biclustering on the source-vs-target influence matrix (panel A) isolates a set of spatial tiles that are, on average, good predictors for all other tiles. Using this set, we use a Voronoi decomposition of the city (Panel B), which realizes an automatic spatial decomposition of the urban space, driven by event predictability.