ORIGINAL ARTICLE

Measurement of repeat effects in Chicago's criminal social network

CrossMark

Paul Kump [a,*], **David Haro Alonso** [a], **Yongyi Yang** [a], **Joseph Candella** [b], **Jonathan Lewin** [b], **Miles N. Wernick** [a]

[a] *Medical Imaging Research Center and Dept. of Electrical and Computer Engineering, Illinois Institute of Technology, 3300 S. Federal St., Chicago, IL 60616, USA*
[b] *Chicago Police Department, 3510 S. Michigan Ave., Chicago, IL 60653, USA*

Abstract The "near-repeat" effect is a well-known criminological phenomenon in which the occurrence of a crime incident gives rise to a temporary elevation of crime risk within close physical proximity to an initial incident. Adopting a social network perspective, we instead define a near repeat in terms of geodesic distance within a criminal social network, rather than spatial distance. Specifically, we report a statistical analysis of repeat effects in arrest data for Chicago during the years 2003–2012. We divide the arrest data into two sets (violent crimes and other crimes) and, for each set, we compare the distributions of time intervals between repeat incidents to theoretical distributions in which repeat incidents occur only by chance. We first consider the case of the same arrestee participating in repeat incidents ("exact repeats") and then extend the analysis to evaluate repeat risks of those arrestees near one another in the social network. We observe repeat effects that diminish as a function of geodesic distance and time interval, and we estimate typical time scales for repeat crimes in Chicago.

1. Introduction

* Corresponding author at: 1820 W. Fletcher, Apt. 2, Chicago, IL 60657, USA. Tel.: +1 815 715 7174.
E-mail addresses: pkump@iit.edu (P. Kump), dharoalo@hawk.iit.edu (D.H. Alonso), yangyo@iit.edu (Y. Yang), Joseph.Candella@chicagopolice.org (J. Candella), jonathan.lewin@chicagopolice.org (J. Lewin), wernick@iit.edu (M.N. Wernick).

Criminological studies have shown that crime is not uniformly distributed among victims, arrestees and places, with repeat crimes playing a fundamental role [5]. In fact, about half of all crimes in the United States are committed by repeat arrestees [19]. Moreover, some reports [4] have suggested a high degree of overlap between victim and arrestee populations, and research has already demonstrated that victims of personal or property crimes and of gun violence experience elevated crime risks within months of an instigating incident [8,16]. Thus, it is reasonable to expect the same effect to exist

among the arrestee population, which has implications for the prevention of crime [10,21].

In the field of criminology there is a well-known phenomenon known as the *near-repeat effect*, which refers to a tendency for crime risk to be temporarily increased within the near vicinity of recent crime incidents (i.e., incidents that have taken place nearby in both space and time; e.g., [10,20]).

In this paper, we hypothesize the existence of a different kind of near-repeat effect, in which we modify the definition of "near" to refer not to geographical relationships among crime incidents, but rather to crime-related interpersonal relationships among the individuals involved in these incidents. We measure the notion of interpersonal distance by using the well-established concept of "degrees of separation," which is known technically as *geodesic distance*. In this paper we define geodesic distance in a criminal social network as follows: If Person A and Person B have been arrested previously in connection with the same crime incident (*co-arrested*), they are said to have one degree of separation. If, in turn, Person C has been co-arrested with Person B, then Person C is said to be separated from Person A by two degrees, and so on. Thus, we hypothesize that a crime incident involving Person A, will temporarily increase the crime risk for individuals such as Persons B and C, who are near Person A in this specific sense that we have defined. Later we show that the hypothesized effect does indeed exist within crime data for the city of Chicago.

The immediate motivation for our study is to inform an ongoing collaborative effort between the Chicago Police Department and our research team at the Illinois Institute of Technology. In this initiative, we have successfully developed and deployed prediction algorithms that estimate the risk of future violence for persons with extensive criminal records. We anticipate that our prediction algorithms will be enhanced by exploiting the probabilistic relationship, if it exists, that would be implied by the aforementioned interpersonal near-repeat effect. Thus, the present work not only informs our fundamental understanding of crime behavior, but is also expected to have practical implications for the prediction of crime, which is a rapidly emerging field [18]. We next review basic concepts of the near-repeat effect and social networks so as to place our work in context.

Repeat crimes may occur for various reasons, including event dependence linked to the psyche, actions, and environment of arrestees [10], with some facet of the arrestee's previous experience increasing the chances of participation in a subsequent incident, with the possibility of the effect spreading to others in the same environment and social groups as the original arrestee [16]. With victims and arrestees tending to belong to similar populations, event dependence suggests the formation of crime patterns establishing positive feedback, eventually escalating into situations involving an excess of dangerous persons or groups, and areas of high crime density. Effective policing strategies would identify these situations as they are forming, thus ceasing the spread of further crime.

Repeat crimes can be characterized as either *exact-repeat* or *near-repeat*, depending on whether consecutive incidents occur at the same location, or at a nearby location. [10] considered the spread of repeat effects to neighboring locations by using Monte Carlo methods to determine the likelihood of the observed patterns occurring if no correlation between spatial and temporal distributions exists. Similarly, Ratcliffe and Rengert [20] investigated repeat effects in shootings in Philadelphia, PA, using a modification of a standard Knox test [12]. Short et al. [22] showed that repeat effects exist among burglaries in Long Beach, CA, and further showed that these effects decrease with distance in space and time.

Social networks have recently been used to investigate the influence that an individual has on his peers [16,17]. There are many reasons to suspect the spreading of repeat effects in a criminal social network. First, many violent crimes are driven by emotions created by social relationships and thus occur between persons who know one another [9]. Second, conditions favoring the participation in crime incidents are spread through peer influence [7]. Third, physical objects such as drugs or weapons are usually dispersed through interpersonal connections, implying that the illegal selling and use of these objects also occur through these connections [6].

For the remainder of this paper, we will present repeat analyses on crimes that occurred in Chicago, IL, during the years 2003–2012, by integrating statistical techniques with a social network perspective. Unlike previous repeat studies that have been concerned with spatial locations and geographical distances of crime incidents, we will focus on arrestees and relationships among arrestees. Each incident in our dataset includes a unique identification number for each arrestee taking part in a crime incident, the date of the incident and the type of crime. Of particular interest to police is the behavior of violent criminals [4], so we divide our data into two mutually exclusive datasets – one containing only violent crimes and the other containing all other crimes – and we perform analyses on the two datasets separately. We refer to these datasets as the *violent* dataset and the *non-violent* dataset, respectively. The former consists of 6630 arrestees, while the latter consists of 941,029 arrestees.

Through the construction and analysis of social networks and application of statistical techniques, this paper aims to discover patterns by repeat arrestees in Chicago. In Section 1 we use a Poisson model to describe the situation in which events are independent of one another and occur only by chance, referring to this as the null model. We describe and apply our counting technique to test our datasets against the null model, and show that Chicago's exact-repeat incidents are not due to chance alone. We describe the social network in Section 2, where we measure the spread of repeat effects through the network by keeping track of the geodesic distance between incidents – that is, the network separation of the two arrestees – and applying well-established spatiotemporal descriptive statistics. We show that repeat effects in our data diminish with time and geodesic distance. We conclude the paper with a summary and discussion.

2. Exact-repeat effects

If there were no repeat effects, an individual's participation in a crime incident would be statistically independent of participation in other crime incidents. This model is contradictory to the existence of repeat effects and will serve as our null hypothesis. Moreover, an arrestee cannot participate in simultaneous criminal incidents, as such incidents would simply be thought of as a single event. Independence of the incidents implies a Poisson process [21], as reviewed next.

2.1. The null hypothesis

In a homogeneous Poisson process the probability that k criminal incidents occur within a time interval from t to Δt is

$$P_k(\Delta t) = \frac{(\lambda \Delta t)^k e^{-\lambda \Delta t}}{k!}, \tag{1}$$

where λ is the rate parameter.

It is easy to show that, for this distribution, the wait time τ between incidents obeys an exponential probability law, i.e.,

$$P(\tau) = \lambda e^{-\lambda \tau}. \tag{2}$$

Therefore, if the null hypothesis is correct, and crime incidents are indeed statistically independent, then the time intervals between incidents in which an arrestee participates should be exponentially distributed, with shorter times occurring more frequently, in accordance with Eq. (2).

2.2. Testing for event dependence

In this section we describe the methodologies used to test for event dependence. For our finite-length data, we employed a moving-window counting method [22] that produces a distribution of the observed wait time τ, which can be compared to the null distribution in Eq. (2) to test for the presence of event dependence. We applied this method to our *violent* and *non-violent* datasets separately.

2.2.1. The moving-window counting method

In the moving-window technique, a time window of width τ_{max} weeks is defined, beginning at the time of the first crime incident in the dataset. If one or more arrestees in this initial incident participate in a secondary incident within the time window, the time interval (wait time) τ between the initial and secondary incidents is recorded. The time window then shifts to the next crime incident in the dataset, and the wait time between that incident and the following one is again recorded. This process continues until the entire data record has been analyzed. The wait time observations obtained at the various time-window positions are assigned to bins to form a histogram of counts, and normalized to obtain a probability distribution, which can be compared to the null distribution in Eq. (2). Note that, in general, the number of observations is greater than the number of repeat arrestees because not all arrestees will participate in a second incident. Hence, order-one arrestees (i.e., arrestees who are arrested only once in the study period) influence the probabilities not by contributing time intervals τ, but by increasing the count. Incidents occurring within the final τ_{max} weeks of the data record – the so-called "buffer period" – are not counted and do not contribute to the probability distribution, because any incident occurring during that period will have its associated time window falling beyond the end of the data record and thus no data can be observed for such incidents.

The exact choice of τ_{max} is not critical, but extreme choices will not allow sufficient data for the opportunity to disprove the null hypothesis. One extreme is to choose τ_{max} as the length of the study's time frame. In this case, all of the data fall within the buffer period and are discarded. The other extreme, $\tau_{max} = 0$, only allows for measurements of $\tau = 0$ nothing larger. For our study, we choose $\tau_{max} = 104$ weeks (two years,

or one-fifth of the study's time frame) corresponding to a buffer period of the years 2011–2012.

The counting scheme with $\tau_{max} = 104$ weeks, applied to a single arrestee, is illustrated by way of an example timeline in Fig. 1, in Supplementary material. The first incident occurs 212 weeks into the study period, and an observation is conducted to determine whether a repeat incident occurs within the following 104 weeks. In this example, a repeat incident has indeed occurred, so the time interval $\tau = 96$ $(308 - 212 = 96)$ is recorded, and the analysis continues through the timeline. The next incident to occur is the one following 308 weeks; this was the repeat incident for the previous measurement and is now a potential instigating incident. Again, an observation is made, but this time a repeat incident does not occur within the time window of $\tau_{max} = 104$ weeks. The analysis continues up to the incident occurring 413 weeks into the study, and it is observed that two later incidents fall within the time window. Since we are counting inter-arrival times of a possible Poisson process, we use the first of these two events, and the time interval $\tau = 7$ $(420 - 413 = 7)$ is added to the record. The buffer period begins after week 416; any possible instigating incident during this period cannot be checked for secondary incidents over the entire time window, and so no further observations are made.

The counting scheme in the preceding example is repeated incident by incident for every arrestee in the dataset. The records of τ are combined into one master record, and the numbers of observations are summed across all arrestees, yielding the total number of observations N_0. We applied this method to the *violent* and *non-violent* datasets separately; a graph of the respective histograms is shown in Fig. 2, in Supplementary material. For the *violent* dataset, we made a total of $N_0 = 4975$ observations, 398 of which were followed by a repeat incident in the following 104 weeks. $N_0 = 261,006$ observations were made for the *non-violent* dataset, and 130,503 observations were followed by a repeat incident within the 104-week time window. Both histograms exhibit more short time intervals than long ones.

2.2.2. Observed data versus the null hypothesis

First consider the *violent* dataset and the upper histogram in Fig. 2, containing $N_0 = 4975$ observations. Normalizing the histogram to obtain a probability distribution yields the solid blue line in the upper graph of Fig. 3, in Supplementary material. By summing the probabilities, we see that about 8% of all violent crimes are followed by an exact repeat occurring within 104 weeks of the instigating incident. Next, we determined the least-squares fit of Eq. (2) to the observed data to estimate the rate parameter. For the *violent* dataset, the least-squares estimate of λ was found to be $\hat{\lambda} = 6.989 \times 10^{-4}$. The function

$$p(\tau) = \hat{\lambda} e^{-\hat{\lambda}\tau}$$

is shown as the dashed red curve in the upper graph of Fig. 3. We applied the same procedure to the *non-violent* histogram with $N_0 = 261,006$ observations. Its probability distribution is shown in the bottom of Fig. 3 along with its best-fit curve, determined to have $\hat{\lambda} = 6.400 \times 10^{-3}$. Again, summation of the probabilities reveals that 50% of non-violent crimes are followed by an exact-repeat incident occurring within 104 weeks of the instigating incident. As we will establish quantitatively

in the next section, the null hypothesis distribution poorly describes the observed data for small values of τ, indicating that event dependence is present in the data.

2.2.3. The Kolmogorov–Smirnov test

To test the validity of the exponential distribution in Eq. (2) as a model of our Chicago repeat data, we use the well-known Kolmogorov–Smirnov (K–S) goodness-of-fit test [14,11], the basic premise of which is to compare the observed cumulative distribution function (CDF) with the CDF of the assumed null distribution. If the maximum discrepancy between the observed and null CDFs is larger than a specific confidence threshold for a given sample size, the null distribution is deemed an unacceptable model of the data. On the other hand, if the discrepancy is less than the threshold, the null distribution cannot be rejected.

Let $F(\tau)$ be the CDF of the null distribution in Eq. (2), that is

$$F(\tau) = \sum_{t=0}^{\tau} p(t),$$

and similarly let $\widehat{F}(\tau)$ be the CDF of the observed data. The test statistic of the K–S test is the maximum absolute difference between $F(\tau)$ and $\widehat{F}(\tau)$. Let this maximum difference be denoted by

$$D_{N_0} = \max_{\tau} |\widehat{F}(\tau) - F(\tau)|.$$

K–S test of size α compares D_{N_0} with the confidence threshold value $D_{N_0}^{\alpha}$ defined by

$$\Pr(D_{N_0} \leqslant D_{N_0}^{\alpha}) = 1 - \alpha,$$

where N_0 is the number of degrees of freedom (histogram normalization factor).

If the observed value of D_{N_0} is greater than the threshold $D_{N_0}^{\alpha}$, the null hypothesis can be rejected at the $(1 - \alpha) \times 100\%$ confidence level. In our study, we chose $\alpha = 0.05$ corresponding to a 95% confidence interval. The degrees of freedom for our *violent* and *non-violent* datasets were $N_0 = 4975$ and $N_0 = 261,006$, respectively. Significance thresholds for $\alpha = 0.05$ and for $N_0 > 50$ are given in Ang and Tang [1] as $D_{N_0}^{\alpha} = 1.36/\sqrt{N_0}$, which yields $D_{N_0}^{\alpha} = 0.019$ and $D_{N_0}^{\alpha} = 0.003$ for our two datasets.

Fig. 4, found in Supplementary material, shows, for each dataset, the observed CDF, $\widehat{F}(\tau)$, and the null CDF, $F(\tau)$, for $\tau \leqslant 104$, as well as the maximum difference D_{N_0} between the two. Our *violent* dataset yielded a K–S test statistic of 0.0251, which is greater than the confidence threshold of 0.019, and so we can reject the null hypothesis that violent incidents are statistically independent. We can draw a similar conclusion with our *non-violent* dataset, which produced a K–S test statistic of 0.016, which is greater than the threshold of 0.003. It should be noted that one cannot infer conclusions about the precise value of τ for which D_{N_0} occurs (which may not even be unique); the test can only reject the null hypothesis and does not provide insight as to the time frame of elevated repeat risk. However, finding a τ for which $|\widehat{F}(\tau) - F(\tau)| > D_{N_0}^{\alpha}$ means, at the very least, that a sufficiently large time window τ_{max} was chosen.

3. Social network repeat effects

Social networks are graph structures composed of nodes (in our case, arrestees), and edges, which form links between these nodes. Social networks can help explain observed local and global patterns within these structures, and identify nodes from which other nodes draw influence.

In this study each arrestee in a dataset defines a node having a unique label within the social network. For purposes of anonymization, we refer to each arrestee using a unique internal record (IR) number. When two or more subjects are arrested as part of the same incident, their respective nodes become connected in the social network. The edge connecting this pair of nodes forms on the day the incident occurred; thus, the social network is constantly evolving. The network is an undirected graph, as there is an inherent symmetry between co-arrestees: if IR1 is arrested with IR2, then IR2 is arrested with IR1. For simplicity in this initial investigation, the edges are unweighted (two subjects are either connected or not, and no connection strength is assigned to the edge that connects them). Although our network is dynamic, unweighted, and undirected, the analysis to follow can readily be adapted to other types of networks.

Although we will separately analyze the repeat behavior of violent and non-violent crimes, we use a single, common social network for all arrestees. We measure the separation between two subjects in the social network using the geodesic distance between their nodes, i.e., the minimum number of edges that must be traversed to move between the nodes. It should be noted that the network contains isolated cliques, i.e., subsets of nodes having edges that do not connect to any node outside the subset. Our network contains nearly three million nodes and 900,000 edges. The largest geodesic distance between any two connected nodes is 37, the largest isolated clique contains 474 nodes, and the smallest isolated cliques consist of just one node.

3.1. Near-repeat counting method

Our process of measuring near-repeat time intervals is a complicated one that is best explained by way of an example. Fig. 5, found in Supplementary material, shows the incident timelines of two subjects, IR1 and IR2, who become immediate neighbors in the network after becoming linked by a co-arrest 100 weeks into the time period studied. In our study, we used the first four years (208 weeks) of data as an initial network setup, or "burn-in" period, during which the network evolves without taking measurements. This is to ensure that the measurements are taken from a network in steady state. In the example in Fig. 5, after the burn-in period, IR1 is arrested in week 209, and his immediate neighbor in the network, IR2, is arrested in week 245. Thus we obtain a measurement of the wait time between arrests of $\tau = 36$ $(245 - 209 = 36)$, indicated in Fig. 5 with a blue line connecting the two incidents. The next arrest occurs in week 212, again involving IR1, after which the soonest arrest of IR2 is again in week 245. Therefore the next measurement of wait time is $\tau = 33$ $(245 - 212 = 33)$. Now the arrest of IR2 in week 245 becomes the instigating incident for the arrest of IR1 in week 275, and $\tau = 30$ is recorded $(275 - 245 = 30)$. Both of the subjects are arrested in week 275, and in this scenario, we count

each arrest as being the instigating incident for the other and hence, $\tau = 0$ is recorded twice. The counting procedure continues in this manner.

An absence of network data for IR1 and IR2 in week 275 indicates that the two arrests are related to separate crime incidents (as is the case for the two arrests in week 201). This is significant because we keep track of the elapsed time since the two were linked and allow the link to expire if this time exceeds four years. By removing connections that become "stale" after several years, we maintain a network with relevant relationships. A link may be renewed should the subjects be arrested together again. In the present example, IR1 and IR2 are linked in week 100 and the link is never renewed. Thus, their link expires in week 308, shown as a vertical dashed line in the figure, and no measurements are taken beyond that time. By choosing an expiration time equal in duration to that of the network burn-in time, the network temporal window is held constant in duration; it is never longer or shorter than four years.

This example illustrates the counting method for arrestees separated by a geodesic distance of $n = 1$ in the social network, but the method is similar for greater values of n. Algorithm 1 in Supplementary material gives the pseudocode algorithm to obtain the wait time distribution for any $n \geqslant 1$.

In a near-repeat analysis (in contrast to our exact-repeat analysis), we would not expect the time intervals to be exponentially distributed should the null hypothesis of event independence be true. This is because the arrests of two subjects in a single incident violate the requirement of non-simultaneity. If independent events are allowed to occur simultaneously, then the time intervals between events will not be exponentially distributed [21]. Nevertheless, statistical testing of our observed distributions can be done using Monte Carlo simulations to estimate a null distribution.

3.2. Testing for elevated clustering in time and geodesic distance

The goal of this section is to determine whether the observed near-repeat distributions for different geodesic distances $p_n(\tau)$ exhibit elevated clustering when compared to the null hypothesis of no event dependence across time and geodesic distance. Originally developed to detect only spatial clustering, the O-ring statistic [13,23] and Ripley's K [23] are spatial descriptive statistics, with roots in epidemiology, that can be extended to detect clustering in a spatiotemporal sense [2,15]. We calculated both statistics using our observed distributions, and in order to compare the statistics to those under the null hypothesis, we performed Monte Carlo simulations in which the dates of the subject-date 2-tuples are randomly permuted. Reshuffling the date data is a common strategy in this situation when the null distribution is unknown. Originally applied in Besag and Diggle [3], this strategy was also implemented in Ratcliffe and Rengert [20] to evaluate near-repeat shooting patterns in the city of Philadelphia, PA.

Let τ_n^{ij} be the time interval between incidents i and j given that the arrestees are separated by geodesic distance n. The O-ring statistic for distance n is given by the following:

$$O_n(\tau) = \frac{1}{\tilde{\lambda}N} \sum_{i \neq j} I(\tau_n^{ij} = \tau), \qquad (3)$$

where N is the total number of incidents, $\tilde{\lambda}$ is the so-called "point density," and $I(\cdot)$ is the indicator function

$$I(a = b) = \begin{cases} 1, & a = b, \\ 0, & a \neq b. \end{cases}$$

In the traditional spatial case, the point density $\tilde{\lambda}$ is usually chosen as the average number of points per unit area. Doing so gives the researcher a meaningful and convenient way to quantify the spatial clustering relative to that expected by a Poisson process. However, as explained earlier, there is no reason to believe our data follow a Poisson process; however, analogous to the spatial case, we can still define $\tilde{\lambda}$ as a rate estimated as N divided by the total number of subjects divided by the total number of weeks in the study.

In the context of near repeats, the summation in Eq. (3) is simply a count of the number of pairs of arrests occurring exactly τ weeks apart involving subjects separated by geodesic distance n. We have already performed this count; it is $N_n \cdot p_n(\tau)$. Therefore, we can express the O-ring statistic in Eq. (3) as

$$O_n(\tau) = \frac{1}{\tilde{\lambda}N} N_n \cdot p_n(\tau). \qquad (4)$$

The closely related Ripley's K function is a cumulative count of time intervals and can be expressed similarly as

$$K_n(\tau) = \frac{1}{\tilde{\lambda}N} \sum_{i \neq j} I(\tau_n^{ij} \leqslant \tau) = \sum_{t=0}^{\tau} O_n(t). \qquad (5)$$

The K and O-ring statistics, while similar, answer slightly different questions. For example, if we hypothesize that near-repeat effects only last for a certain duration and then "expire," and the accumulating nature of the K statistic should identify the time at which the effect expires. In this case, we would expect elevated K statistics at short time intervals and moderate K statistics falling within confidence intervals at larger time intervals. The O-ring is more of an instantaneous metric identifying specific times of heightened near-repeat risk. An elevated O-ring statistic at, say, $\tau = 4$ weeks would imply clustering of near-repeat incidents four weeks after instigating incidents and does not confound with the O-ring statistics at larger values of τ. Both statistics are of interest to us; therefore we consider both in our analysis.

We first consider our *violent* dataset, and plot the observed O-ring and K statistics for $n = 1, 2, 3$ and $0 \leqslant \tau \leqslant 208$ in Fig. 6, in Supplementary material. For comparison to the null hypothesis, we randomly permute the dates of all arrests while holding the subjects fixed, and, with the new data, calculate and plot the O-ring and K statistics. This process is repeated for 95 random permutations for the sake of constructing a 95% confidence envelope, also plotted in Fig. 6. Randomly permuting the dates has the effect of observing how O-ring and K statistics would appear should there be no spatiotemporal relationships [20].

We observe from Fig. 6 elevated O-ring and K statistics within the first 25 weeks for the cases of $n = 1, 2$ neighbors, with the effect extending past 80 weeks for the K statistic. For the $n = 3$ case, the figure shows that the statistics are completely within the envelope of what can be expected under the null hypothesis and are therefore not statistically significant at the 95% confidence level. Not shown are the statistics for $n = 4$ neighbors for which we also found no significant dependence.

Next we consider the *non-violent* dataset for which we conducted the same analysis as for the *violent* dataset. The results are shown in Fig. 7, in Supplementary material. For $n = 1, 2$, the figure shows that the O-ring statistic falls outside of the 95% confidence envelope for small values of τ during which the statistic exceeds the upper limit of the confidence interval. For larger n, the O-ring statistic falls entirely within the confidence envelope and is therefore not statistically significant at those values of n. The figure also shows an elevated K statistic for $n = 1, 2, 3, 4$ and a K statistic values that are not significant for $n = 5$. We repeated the analysis for $n = 6$ and $n = 7$ and found that neither the O-ring nor the K statistic showed significant event-dependence effects in either case.

3.3. Discussion

The above pioneering approach of fusing the O-ring and K statistics with a social network analysis provides knowledge of the temporary influence one has on his peers within the network. This influence may be described by the strength of the influence – that is, the maximum geodesic distance for which the influence is observed – and its persistence, i.e., for how long, temporally, the influence is observed.

Since the O-ring statistic is, as explained previously, a measure of instantaneous effect, it provides insight concerning the time scales of repeat incidents in our Chicago datasets. For instance, the O-ring statistic is elevated for our *violent* dataset during roughly the first 25 weeks for $n = 1$, indicating (possibly for the reasons outlined in the introduction) an increased chance of a violent repeat crime occurring within about six months of an initial violent crime for subjects directly connected in the network. During this time, we observe 58% more repeat incidents than expected by chance. The K statistic reaches a maximum separation from its confidence envelope during this time period before it begins to approach and finally penetrates the envelope at about 90 weeks. Thus the total number of violent repeat crimes separated by less than 90 weeks is more than what would be expected by chance alone.

A similar argument can be made for the violent $n = 2$ case in which the O-ring statistic in Fig. 6 implies an elevated probability of violent repeat crimes occurring within just two months of an initial violent crime. During this time, we observe 70% more repeat incidents than expected by chance. The K statistic demonstrates an elevation in the total number of violent repeat crimes separated by less than 80 weeks. Finally, our results for $n = 3$ and $n = 4$ ($n = 4$ not shown in the figure) demonstrate no statistically significant violent near-repeat effects; thus the concept of "near" for violent crimes may be limited to arrested individuals within two degrees of network separation.

For the *non-violent* dataset, the O-ring statistic falls outside of the confidence envelope for neighbors of distances $n = 1$ and $n = 2$, and is entirely contained within the envelope for larger n. For $n = 1$, the elevated O-ring statistic at $\tau = 0$ and $\tau = 1$ demonstrates an increased risk during the first and second weeks following an instigating incident. In fact, there are 28% more near-repeat incidents in the first two weeks in the dataset than is predicted by the null hypothesis. Similarly, for $n = 2$ neighbors, the elevated statistic at $\tau = 0$ implies an increased risk during the first week following an instigating

incident. There are 17% more near-repeat incidents in the first week than predicted by the null hypothesis.

The K statistics in Fig. 7 show a near-repeat effect which decreases in both time and network space. For $n = 1$ neighbors, more near-repeat crimes are observed within 22 weeks of instigating incidents than would be expected by chance; however, this time frame decreases as subjects become further separated. For example, there are more repeat crimes than expected only for the first 20 weeks for $n = 2$, 10 weeks for $n = 3$, and 8 weeks for $n = 4$, and the number of crimes are as expected for $n = 5$ and $n = 6$ ($n = 6$ not shown in the figure). During these time frames, we see 10.47% more crimes than expected for $n = 1$, 4.91% for $n = 2$, 3.06% for $n = 3$, and 2.96% for $n = 4$. The K statistics show repeat effects at $n = 3$ and $n = 4$ while the O-ring statistics do not. This is because the two statistics answer slightly different questions. The K statistic finds elevated crimes for the $n = 3$ and $n = 4$ cases accumulated up to 10 weeks and 8 weeks, respectively; however, according to the O-ring statistic, there is no single week during these periods in which we should expect an elevated number of repeat crimes.

4. Conclusion

This work is motivated by our interest in identifying factors that may be predictive of an individual's involvement in future violence to enhance the predictive models we have developed as part of the Chicago Police Department's predictive policing program. Specifically, we have developed a very successful prediction model that evaluates the crime risk of individuals with very severe patterns of involvement in crime.

This paper demonstrates conclusively that an individual's crime risk increases for a period of roughly 25 weeks (about 6 months) following a crime involving a person within two or less degrees of separation from an individual involved in that crime. To make practical use of this finding, we are currently in the process of revising our prediction model by including new variables pertaining to first- and second-degree connections, as well as a weighting function that gives greater emphasis to the effect of recent crime incidents, as opposed to those in the more distant past. In addition to the practical significance for our current project, the findings of this paper provide us with fundamental insights about patterns of crime and its temporal and interpersonal behavior, for both violent and non-violent crime types.

In this work, to test for the presence of a social-network near-repeat effect, we conducted statistical analyses of arrest data to measure exact-repeat and near-repeat effects within Chicago's criminal social network for both violent and non-violent crimes. For exact-repeats, we modeled the situation in which arrests are independent of one another as a Poisson process, and therefore inter-arrival times were compared to an exponential distribution. We selected the distribution's rate parameter to minimize the residual sum of squares and quantify the goodness-of-fit using the popular K–S test. The K–S test rejects the null hypothesis of independence for both of our datasets with 95% confidence.

We measured near-repeat effects by considering how the effects spread through a social network. Each node in the network is a unique arrestee, and two nodes become connected should two subjects be arrested as part of the same unlawful

incident. With the network in place, we recorded the time intervals between repeat incidents, this time keeping track of "distances", or the degree separation between graph nodes. By recording these geodesic distances, we allowed for the separate examination of temporal and spatial effects using Ripley's K and O-ring statistics. In order to compare these statistics to those of a null distribution of no dependence among incidents (across time and geodesic distance), we randomly permuted arrest dates and then recalculated the test statistics. The results showed elevated repeat effects that decrease over time and geodesic distance within the network.

For violent crimes in particular, we have observed that the near-repeat effect extends to neighbors separated by a geodesic network distance of $n = 2$, with the effect most prominent for neighbors of $n = 1$. Further, we have seen that this effect expires at most 25 weeks after an instigating crime. Therefore, when considering the role one individual's crime plays on the risk of a second individual, we must only concern ourselves with crimes from $n = 1$ and $n = 2$ neighbors, from at most 25 weeks ago. As explained earlier, these findings will inform future predictive models for identifying individuals at high risk for future violence.

Acknowledgments

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.aci.2016.01.002.

References

[1] A. Ang, W. Tang, Probability Concepts in Engineering: Emphasis on Application to Civil and Environmental Engineering, Wiley, 2006.

[2] M.-L. Benot, A.-K. Bittebiere, A. Ernoult, B. Clement, C. Mony, Fine-scale spatial patterns in grassland communities depend on species clonal dispersal ability and interactions with neighbours, J. Ecol. 101 (2013) 626–636.

[3] J. Besag, P.J. Diggle, Simple Monte Carlo tests for spatial pattern, Appl. Stat. 26 (1977) 327–333.

[4] A.A. Braga, Gun Violence Among Serious Young Offenders, Technical Report, U.S. Department of Justice, 2008.

[5] M. Comeau, Repeat and Near-Repeat Burglary Victimization in Rochester, NY, Technical Report, Center for Public Safety Initiatives, 2014.

[6] P.J. Cook, J. Ludwig, S.A. Venkatesh, A.A. Braga, Underground gun markets, Econ. J. 117 (2007) 558–588.

[7] J. Fagan, D.L. Wilinson, Guns, youth violence, and social identify in inner cities, Crime Justice 24 (1998) 105–188.

[8] G. Farrell, Preventing repeat victimization, Crime Justice (1995) 469–534.

[9] R.B. Felson, H.J. Steadman, Situational factors in disputes leading to criminal violence, Criminology 21 (1983) 59–74.

[10] S.D. Johnson, Repeat burglary victimization: a tale of two theories, J. Exp. Criminol. 4 (2008) 215–240.

[11] A. Kolmogorov, Sulla determinazione empirica di una legge di distribuzione, Giornale Dell' Instuto Italiano degli Attuari 4 (1933) 83–91.

[12] M. Kulldorff, The Knox method and other tests for space-time interactions, Biometrics 55 (1999) 544–552.

[13] R. Law, J. Illian, D. Burslem, G. Gratzer, C.V.S. Gunatilleke, I. Gunatilleke, Ecological information from spatial patterns of plants: insights from point process theory, J. Ecol. 97 (2009) 616–628.

[14] S.C. Loftus, L.L. House, M.C. Hughey, J.B. Walke, M.H. Becker, L.K. Belden, Dimension Reduction for Multinomial Models Via a Kolmogorov-Smirnov Measure (KSM), Technical Report, No. 15-1, Virginia Tech, 2015.

[15] R.J.Q. McNally, J. Rankin, M.D.F. Shirley, S.P. Rushton, T. Pless-Mulloli, Space-time analysis of down syndrome: results consistent with transient pre-disposing contagious agent, Int. J. Epidemiol. 37 (2008) 169–1179.

[16] A.V. Papachristos, A.A. Braga, D.M. Hureau, Social networks and the risk of gunshot injury, J. Urban Health 89 (2012) 992–1003.

[17] A.V. Papachristos, D.M. Hureau, A.A. Braga, The corner and the crew: the influence of geography and social networks on gang violence, Am. Sociol. Rev. 78 (2013) 417–447.

[18] W.L. Perry, B. McInnis, C.C. Price, S.C. Smith, J.S. Hollywood, Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations, RAND Corporation, 2013.

[19] Pew Center on the States, State of Recidivism: The Revolving Door of America's Prisons, 2011.

[20] J.H. Ratcliffe, G.F. Rengert, Near-repeat patterns in Philadelphia shootings, Secur. J. 21 (2008) 58–76.

[21] Y.A. Rozanov, Probability Theory: A Concise Course, Dover, 1977.

[22] M.B. Short, M.R. D'Orsogna, P.J. Brantingham, G.E. Tita, Measuring and modeling repeat and near-repeat burglary effects, J. Quant. Criminol. 25 (2009) 325–339.

[23] T. Wiegand, M. Andersson, P. Krieger, Spatial point pattern analysis of neurons using Ripley's K-function in 3D, Front Neuroinform. 4 (2010).