IOP PUBLISHING INVERSE PROBLEMS

Inverse Problems **27** (2011) 115013 (15pp) doi:10.1088/0266-5611/27/11/115013

Reconstruction of missing data in social networks based on temporal patterns of interactions

Alexey Stomakhin, Martin B Short and Andrea L Bertozzi

Department of Mathematics, University of California, Los Angeles, CA, USA

E-mail: alexey@math.ucla.edu, mbshort@math.ucla.edu and bertozzi@math.ucla.edu

Received 11 March 2011, in final form 30 September 2011
Published 28 October 2011
Online at stacks.iop.org/IP/27/115013

Abstract

We discuss a mathematical framework based on a self-exciting point process aimed at analyzing temporal patterns in the series of interaction events between agents in a social network. We then develop a reconstruction model that allows one to predict the unknown participants in a portion of those events. Finally, we apply our results to the Los Angeles gang network.

(Some figures may appear in colour only in the online journal)

1. Introduction

Prediction of missing information is an important part of data analysis in social sciences [1–3]. The examples studied in the literature, mostly by statisticians, include reconstruction of the unknown connections in a social network [4, 5], analyzing non-ignorable non-responses in a survey sampling [6, 7] and many others. The most common way to deal with missing values is to replace them by some plausible estimates using known or model-based cross-dependencies over the network in question.

However, these methods do not typically consider networks that change with time, when another source of information is given by the temporal patterns arising from the network evolution. Such networks are the primary object of study in this paper. As our main example, we consider the gang rivalry network in the Los Angeles policing district Hollenbeck [8]. Police data on gang crimes from 1999 to 2002 reveal temporal clustering of gang interaction events, which is demonstrated in figure 1. These temporal patterns can be used to solve the following inverse problem: predict the participants of the gang-related crimes if some of them are not known.

For a given pair of agents, the interaction events can either be independent, following a Poisson process, or temporally dependent, in which case the occurrence of one event can change the likelihood of subsequent events in the future. Such event dependence for the Los Angeles gang network has been established in [9], where a Hawkes process [10, 11], commonly used in seismology to model earthquakes [12, 13], was compared to inter-gang violent crimes.

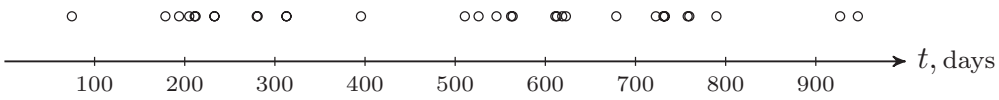

Figure 1. Temporal clustering of the interaction events between Clover and East Lake gangs in Los Angeles, during the period 1999–2002.

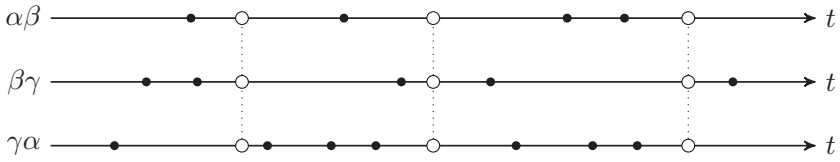

Figure 2. Graphical representation of the problem.

This paper is organized as follows. In sections 2 and 3, we formalize the problem and describe a model of interaction between network agents based on a Hawkes process. In section 4, we propose a way of predicting the unknown participants of interaction events, which we formulate as a constrained optimization problem. In sections 5 and 6, we analyze our method and the solution it gives. Finally, in section 7 we present and discuss the prediction results.

2. Problem formulation

We model a social network as a graph with nodes representing the agents and edges, or binary links [5], indicating whether or not the corresponding pair of agents interact. We further look at the series of pairwise interaction events between the agents, characterized by their occurrence times and the pairs involved. We assume that the network structure represented by the graph does not change with time, although each pair of interacting agents can have its own prescribed model of behavior that might involve some time dependence. Suppose all the times of the events are known, but for some of them, data for one or both of the participants are missing. The problem is to reconstruct the missing data about the participants based on the behavioral model.

Before we proceed, let us discuss a convenient graphical representation of the problem shown in figure 2. Here, we deal with a network consisting of three agents α, β and γ, with all pairs being active. The black points correspond to the events without any missing information. All events are ordered in time and there is a separate timeline for each pair of agents. The incomplete events, which are those with missing data about the participants, cannot be assigned to any particular timeline and are therefore represented via vertical series of white circles. Our goal is to replace each vertical set of white circles with a black circle on one of the timelines in a way that will give the most plausible picture in accordance with the model.

Returning to the network of gangs in Los Angeles: there are 29 agents and the binary links indicate the existing rivalries between them, shown in figure 3. In case of a rivalry, we have a series of crimes corresponding to the interaction events. These are typically murders, shots fired, etc. The data capture the information about which two gangs were involved in a crime; however, for a large fraction of them only victim affiliation is provided. The problem in this case is to estimate the affiliation of the unknown offenders.

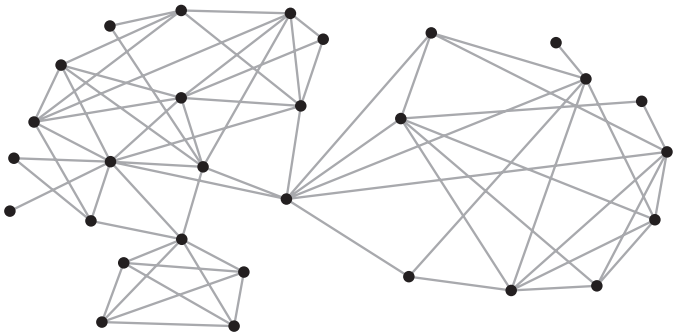

Figure 3. Graph of the gangs network in the Los Angeles policing district Hollenbeck [8]. Each of the 29 gangs is represented by a node, and the edges indicate the presence of rivalries between them.

3. Agent interaction model

A Hawkes process [10, 11] is a self-exciting point process commonly used in seismology to model earthquakes [12, 13] and defined by its intensity function

$$\lambda(t) = \mu + \theta \sum_{t_i < t} g(t - t_i). \tag{1}$$

The intensity function $\lambda(t)$ is partitioned into the sum of a Poisson background rate μ and a self-exciting component, through which events trigger an increase in the intensity of the process. The elevated rate spreads in time according to the kernel g, with θ being the scaling factor of the effect. In other words, each event generates a sequence of offspring or repeat events, which leads to temporal clustering. This agrees with the evidence that retaliations are commonplace among rival gangs [14, 15]. A similar approach was used to model repeat and near-repeat burglary effects in [16, 17] and temporal dynamics of violence in Iraq in [18], where self-excitation is one of the key qualitative features of the process.

We assume that the interaction events for each pair of agents occur independently according to a Hawkes process. This assumption of independence is based on the tentative conclusion in [19]. That is, the network of Hollenbeck gangs may be in a homogeneous state, meaning that gangs are not tightly coupled to one another. Thus, if gang α is fighting with gang β, and gang γ begins attacking α, then α easily switches away from its rivalry with β to begin fighting γ. This switching is largely a random, independent event in the homogeneous state.

We make no exclusions for inactive pairs since for those we simply have $\mu = 0$, and it is also useful to set $\theta = 0$ to avoid confusion in the following analysis. For the function g, as in [9], we use an exponential distribution, which gives

$$\lambda(t) = \mu + \theta \sum_{t_i < t} \omega \, \mathrm{e}^{-\omega(t - t_i)}. \tag{2}$$

Here, ω^{-1} sets the time scale over which the overall rate $\lambda(t)$ returns to its baseline level μ after an event occurs [20]. From the behavioral point of view, θ represents the average number of direct offspring for each event and ω^{-1} is the expected waiting time until an offspring. To

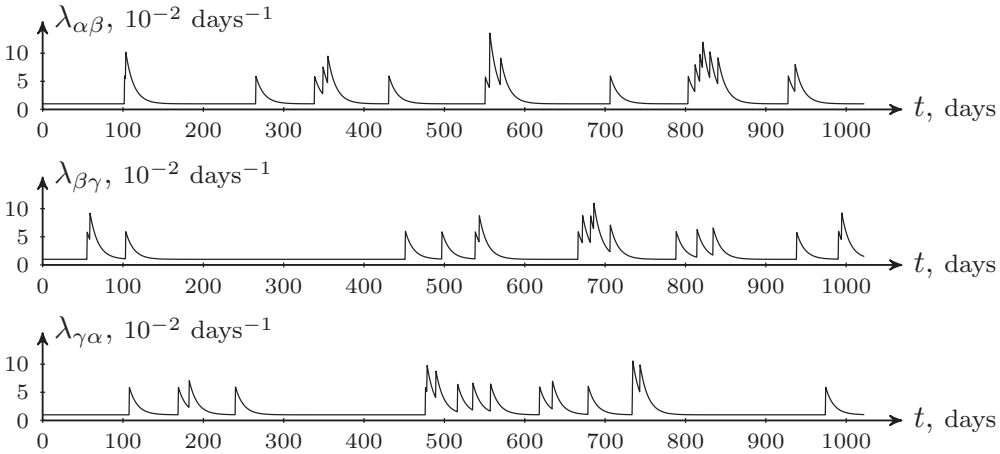

Figure 4. Data generated according to a Hawkes process with the same parameters for each pair of agents: $\mu = 10^{-2}$ days^{-1}, $\omega = 10^{-1}$ days^{-1} and $\theta = 0.5$.

indicate that each pair of agents has its own interaction parameters, we use index notation and write

$$\lambda_{\alpha\beta}(t) = \mu_{\alpha\beta} + \theta_{\alpha\beta} \sum_{t_i^{\alpha\beta} < t} \omega_{\alpha\beta}\, \mathrm{e}^{-\omega_{\alpha\beta}(t - t_i^{\alpha\beta})}, \tag{3}$$

with $\mu_{\alpha\beta}, \theta_{\alpha\beta}, \omega_{\alpha\beta}$ being constants, unique for each pair, and summation over all previous events between the agents α and β. If no confusion is possible, we will omit the indices $\alpha\beta$ to simplify the notations in the future.

In figure 4, we present an example of data generated according to the described model (3) for a network consisting of three agents α, β and γ. Here, the same parameters are used for each pair: $\mu = 10^{-2}$ days^{-1}, $\omega = 10^{-1}$ days^{-1} and $\theta = 0.5$. These have approximately the order of magnitude estimated in [9] for the Los Angeles gang network.

Note here that obtaining the interaction parameters based on the given data is a separate problem which is not addressed in this work. We have some discussion of this in section 8. In what follows, we assume that all the interaction parameters are known and use them to predict missing participants of the incomplete events.

4. Reconstruction method

We will use the following notations:

N	total number of events
n	number of incomplete events
k	number of agents
K	total number of pairs $= k(k-1)/2$.

To solve the prediction problem in question, one could consider the likelihood function, defined on the space of all possible event lists, corresponding to different ways of filling in the missing data, which is to be maximized in order to get the most likely one. Given any *complete* event list, with no missing data, its likelihood is given by (see, for example, [9])

$$\mathcal{L} = \prod_{\alpha\beta} \prod_{t_i^{\alpha\beta}} \lambda_{\alpha\beta}\big(t_i^{\alpha\beta}\big). \tag{4}$$

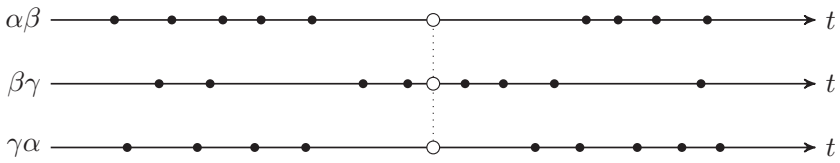

Figure 5. Example of reconstruction based on temporal clustering: agents β and γ are the most likely participants of the incomplete event, as this would place it within a cluster.

The first product is over all possible *unordered* pairs of agents, and the second one is over all events for a fixed pair.

Note that maximizing (4) is a combinatorial type problem since the set of all agent pairs is discrete. One of the possible approaches to this problem would be to use simulated annealing [21] or Monte Carlo Markov chain [22] techniques to estimate the maximum of the likelihood. These methods though being probabilistic metaheuristics usually require problem-dependent tuning, and can be rather slow. Another technique is to consider an approximation to the real likelihood function (4). Some examples of pseudolikelihood methods for point processes were introduced for instance in [23, 24]. Note, however, that no matter what approximation we take, it still will be a function defined over a discrete set. Unfortunately then, there seems to be no significantly more optimal way than 'full search' for solving this problem exactly in the general case, which is very inefficient since its complexity depends exponentially on n.

The goal is therefore to make the maximization problem computationally less expensive, while maintaining the plausibility of the predictions. To do this, one could define a smooth extension of (4) and then look for its maximum, so that some standard continuous optimization method like gradient ascent could be used. This could be achieved by allowing each incomplete event to move continuously between the timelines. However, such an approach is not naturally applicable to (4) due to its multiplicative structure.

We therefore propose the following. We design some reasonable approximation to the real likelihood function (4) such that its continuous extension is physically meaningful. Let us start with the following simple example. Consider a network consisting of three agents α, β and γ with all pairs having the same interaction parameters. Suppose only one event is incomplete and there is no information about its participants. Intuitively, because of the self-exciting nature of the process, the event is less likely to belong to the pairs with no nearby interaction, and more likely to belong to those for which it can be considered as a part of a cluster. For instance, in the situation shown in figure 5, agents β and γ are the most likely participants of the incomplete event, as this would place it within a cluster.

To give this idea a quantitative formulation, we note that clusters correspond to the periods of time with higher values of the intensity functions, which can be seen in figure 4. Hence, for a missing event it would be reasonable to predict the pair with the highest intensity at the moment of the event. It also makes sense from the probabilistic point of view, because given the fact that an event happened at time t the probability of pair $\alpha\beta$ being involved is proportional to $\lambda_{\alpha\beta}(t)$.

Now we construct our energy functional: an approximation to the likelihood function (4) on the space of all possible event lists, corresponding to different ways of filling in the missing data, which is to be maximized in order to get the 'most likely' one. Given any *complete* event list, with no missing data, we define its energy as

$$\Lambda = \sum_{\alpha\beta} \sum_{t_i^{\alpha\beta}} \lambda_{\alpha\beta}\left(t_i^{\alpha\beta}\right). \tag{5}$$

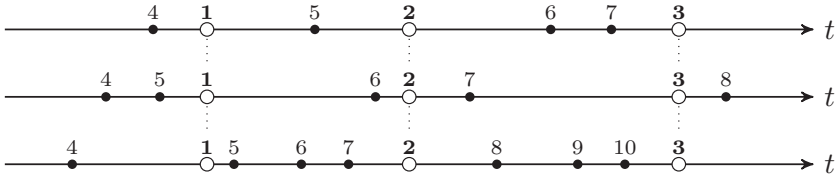

Figure 6. Events enumeration example.

The first summation is over all possible *unordered* pairs of agents, and the second one is over all events for a fixed pair. Here, we basically say that the 'chances' of a pair being involved in an interaction event are equal to its intensity function value at that time. Then, we take the sum over all events. Roughly speaking, the metric defined by (5) assigns higher values to the event lists with denser clusters, which is precisely what we need to get a reasonable prediction. For simplicity of notation, we replace $t_i^{\alpha\beta}$ with i in the summation index, keeping in mind that each pair of agents has their own timeline and system of indices for the events on it. Substituting (3) into (5) gives

$$\Lambda = \sum_{\alpha\beta} \sum_{i,j} \delta_{ij}\mu_{\alpha\beta} + \frac{1}{2}(1-\delta_{ij})\theta_{\alpha\beta}\omega_{\alpha\beta}\, \mathrm{e}^{-\omega_{\alpha\beta}|t_i^{\alpha\beta}-t_j^{\alpha\beta}|}. \tag{6}$$

Thus, Λ is decoupled into the sum of the energies of the events themselves, determined by the background rates, and the sum of the pairwise interaction energies between the events on the same timeline due to self-excitation. Clustering leads to a stronger interaction, increasing the value of Λ. Clearly, functional (6) is invariant with respect to time inversion, which means that each event affects its successors and predecessors in the same way.

As an alternative to (5), one could normalize the intensity functions over all pairs of agents to make them add up to 1, and define the energy functional as

$$\tilde{\Lambda} = \sum_{\alpha\beta} \sum_{t_i^{\alpha\beta}} \frac{\lambda_{\alpha\beta}\left(t_i^{\alpha\beta}\right)}{\sum_{\alpha'\beta'} \lambda_{\alpha'\beta'}\left(t_i^{\alpha\beta}\right)}, \tag{7}$$

an approach that might seem to be more natural from the probabilistic point of view. However, it makes the final optimization problem be solved much more nonlinear and has a drawback discussed in section 5.

Again, maximizing the energy functional (5) or (7) is a combinatorial type problem since the set of all agent pairs is discrete, and there seems to be no significantly more optimal way than 'full search' for solving it in the general case, which is very inefficient. However, unlike the likelihood function (4), it admits a physically meaningful smooth extension. This can be obtained by distributing each of the incomplete events over the timelines with weights that 'add up', in some sense, to 1. Thus, in figure 2, we would replace the white circles with black ones and add weights to each of those; the complete events naturally receive weight 1. We can interpret this to mean that each incomplete event occurred *partially* on every timeline with effect (the jump in the intensity function) proportional to the corresponding weight. This new continuous maximization problem not only gives the most likely participants of an event, but also assigns a weight to each pair showing how likely that pair is to be involved.

To avoid misunderstanding, let us specify how we enumerate the events on a timeline, which does matter now due to the normalization coupling of the pairs. The reader can use figure 6 as a reference. We start with incomplete events and assign them numbers from 1 to n. The order here is not important, as long as it is the same for all timelines. Then, for each

timeline we assign numbers to the complete events starting from $(n+1)$. Thus, there is a separate event indexing system for each timeline, with indices coinciding for the incomplete events.

Using l^2-normalization for the weights, we get the following formulation of the problem:

$$\begin{cases} \max\left\{\sum_{\alpha\beta}\sum_{i,j}\left[\delta_{ij}\mu_{\alpha\beta}m_i^{\alpha\beta}+\frac{1}{2}(1-\delta_{ij})\theta_{\alpha\beta}\omega_{\alpha\beta}\,\mathrm{e}^{-\omega_{\alpha\beta}|t_i^{\alpha\beta}-t_j^{\alpha\beta}|}m_i^{\alpha\beta}m_j^{\alpha\beta}\right]\right\} \\ \sum_{\alpha\beta}\left(m_i^{\alpha\beta}\right)^2=1, \quad \forall\, i=1,\ldots,n \\ m_i^{\alpha\beta}\geqslant 0, \quad \forall\, i=1,\ldots,n, \quad \forall\,\alpha\beta, \end{cases} \tag{8}$$

where $m_i^{\alpha\beta}$ denotes the weight of the event number i on timeline $\alpha\beta$. As we mentioned before, the complete events have weight 1, so $m_i^{\alpha\beta}\equiv 1$ for $i>n$. The objective function is maximized with respect to $m_i^{\alpha\beta}$ for $i\leqslant n$, given the normalization and non-negativity constraints.

One could alternatively choose to use l^1-normalization for the weights, which again might seem to be more natural from the probabilistic point of view. The problem in that case is

$$\begin{cases} \max\left\{\sum_{\alpha\beta}\sum_{i,j}\left[\delta_{ij}\mu_{\alpha\beta}m_i^{\alpha\beta}+\frac{1}{2}(1-\delta_{ij})\theta_{\alpha\beta}\omega_{\alpha\beta}\,\mathrm{e}^{-\omega_{\alpha\beta}|t_i^{\alpha\beta}-t_j^{\alpha\beta}|}m_i^{\alpha\beta}m_j^{\alpha\beta}\right]\right\} \\ \sum_{\alpha\beta}m_i^{\alpha\beta}=1, \quad \forall\, i=1,\ldots,n \\ m_i^{\alpha\beta}\geqslant 0, \quad \forall\, i=1,\ldots,n, \quad \forall\,\alpha\beta. \end{cases} \tag{9}$$

However, this method is unstable with respect to the input data, as we will see in section 5.

Note here that the discrete, combinatorial version of this method can be obtained from (8) or (9) by forcing all weights to be integers

$$\begin{cases} \max\left\{\sum_{\alpha\beta}\sum_{i,j}\left[\delta_{ij}\mu_{\alpha\beta}m_i^{\alpha\beta}+\frac{1}{2}(1-\delta_{ij})\theta_{\alpha\beta}\omega_{\alpha\beta}\,\mathrm{e}^{-\omega_{\alpha\beta}|t_i^{\alpha\beta}-t_j^{\alpha\beta}|}m_i^{\alpha\beta}m_j^{\alpha\beta}\right]\right\} \\ \sum_{\alpha\beta}m_i^{\alpha\beta}=1, \quad \forall\, i=1,\ldots,n \\ m_i^{\alpha\beta}\in\{0,1\}, \quad \forall\, i=1,\ldots,n, \quad \forall\,\alpha\beta. \end{cases} \tag{10}$$

5. Examples

The purpose of this section is to discuss a few examples that will reveal some useful properties of problem (8).

5.1. Example 1: timescale detection

Suppose $N=n=2$, so we have two incomplete events, and suppose we do not have any information at all about the participants. For simplicity, we also assume $\mu_{\alpha\beta}\equiv 0$ and $\theta_{\alpha\beta}\equiv 1$. Then, the problem to be solved according to (8) is

$$\begin{cases} \max\sum_{\alpha\beta}\omega_{\alpha\beta}\,\mathrm{e}^{-\omega_{\alpha\beta}\Delta t}m_1^{\alpha\beta}m_2^{\alpha\beta} \\ \sum_{\alpha\beta}\left(m_i^{\alpha\beta}\right)^2=1, \quad \forall\, i=1,2 \\ m_i^{\alpha\beta}\geqslant 0, \quad \forall\, i=1,2, \quad \forall\,\alpha\beta, \end{cases} \tag{11}$$

with Δt being the time interval between the events. Note that (11) can be written conveniently in vector form as

$$\begin{cases} \max \mathbf{m}_1^T\mathbf{D}\mathbf{m}_2 \\ \|\mathbf{m}_1\|_2=\|\mathbf{m}_2\|_2=1 \\ m_i^{\alpha\beta}\geqslant 0, \quad \forall\, i=1,2, \quad \forall\,\alpha\beta, \end{cases} \tag{12}$$

where we have used the notations

$$\begin{cases} \mathbf{D}=\mathrm{diag}\{\omega_{\alpha\beta}\,\mathrm{e}^{-\omega_{\alpha\beta}\Delta t}\}\in\mathbb{R}^{K\times K} \\ \mathbf{m}_i=\left\{m_i^{\alpha\beta}\right\}\in\mathbb{R}^K,\ i=1,2. \end{cases} \tag{13}$$

From linear algebra, it is well known that the objective function in (12) is maximized when $\mathbf{m}_1 = \mathbf{m}_2 = \mathbf{e}_{\alpha'\beta'}$, the unit vector, such that

$$\alpha'\beta' = \arg\max_{\alpha\beta}\{\omega_{\alpha\beta}\, \mathrm{e}^{-\omega_{\alpha\beta}\Delta t}\}. \tag{14}$$

The maximum of $\omega \mathrm{e}^{-\omega\Delta t}$ is achieved when $\omega = \frac{1}{\Delta t}$. Hence, the solution of problem (11) corresponds to the pair with self-excitation timescale closest to Δt.

Recall that the self-excitation timescale represents the average time until a repeat event occurs. Thus, since all background rates are equal to zero, and therefore the second event must be an offspring of the first one, our method indeed gives the most likely participants. Of course, for prediction purposes the distribution of the weights is not very realistic, because it rules out the possibility of all other pairs being involved. But, as we will see further, there are other mechanisms that make the solution more regularized, which we do not see here due to a specific and, in fact, unrealistic structure of the example. Indeed, this example is in some sense pathological, as there is no way to explain the occurrence of the first event. However, we can think of it as a limiting case when

$$\sum_{\alpha\beta} \mu_{\alpha\beta} \ll \min_{\alpha\beta}\{\omega_{\alpha\beta}\, \mathrm{e}^{-\omega_{\alpha\beta}\Delta t}\}. \tag{15}$$

Then, the first event is a background one, which happened after waiting for a sufficiently long time, and the second one is due to self-excitation, because the probability of it being a background event *from some timeline* is much less than the probability of it being an offspring of the previous event, as follows from (15).

Consider now the alternative energy functional (7) with normalization at each time point, introduced in section 4. Clearly, the maximum value it can achieve, for the example in question, is 1. It happens whenever both events completely belong to the same pair of agents. Thus, this model does not 'see' the dependence of clustering density on self-excitation timescale, and leads to a degenerate solution.

5.2. Example 2: regularization

Suppose $N = n = 1$, so we have only one event which is incomplete, and suppose we do not have any information at all about the participants. Then the problem to be solved according to (8) is

$$\begin{cases} \max \sum_{\alpha\beta} \mu_{\alpha\beta} m^{\alpha\beta} \\ \sum_{\alpha\beta} (m^{\alpha\beta})^2 = 1 \\ m^{\alpha\beta} \geqslant 0, \quad \forall\, \alpha\beta. \end{cases} \tag{16}$$

Problem (16) can be written conveniently in vector form as

$$\begin{cases} \max \boldsymbol{\mu}^T \mathbf{m} \\ \|\mathbf{m}\|_2 = 1 \\ m^{\alpha\beta} \geqslant 0, \quad \forall\, \alpha\beta, \end{cases} \tag{17}$$

where we have used the notations

$$\begin{cases} \boldsymbol{\mu} = \{\mu^{\alpha\beta}\} \in \mathbb{R}^K \\ \mathbf{m} = \{m^{\alpha\beta}\} \in \mathbb{R}^K. \end{cases} \tag{18}$$

The maximizer of (17) is well known from linear algebra to be

$$\mathbf{m} = \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_2}. \tag{19}$$

Thus, the optimal weights, according to our method, are proportional to the corresponding background intensity rates. This is exactly what follows from the probabilistic approach. Indeed, we are dealing with the case where no self-excitation takes place, since there is only one event. Therefore, the probability of a pair being involved in the event is proportional to its background intensity rate.

Consider now the alternative model (9) with l^1-normalization, mentioned in section 4. For this example it gives the following optimization problem:

$$\begin{cases} \max \boldsymbol{\mu}^T \mathbf{m} \\ \|\mathbf{m}\|_1 = 1 \end{cases}. \tag{20}$$

Clearly, the objective function in (20) is maximized when $\mathbf{m} = \mathbf{e}_{\alpha'\beta'}$, the unit vector, such that

$$\alpha'\beta' = \arg\max_{\alpha\beta}\{\mu_{\alpha\beta}\}. \tag{21}$$

We see that the model picks the pair with the highest background rate, and assigns weight 1 to it and 0 to the others. However, this is not the most desirable solution. Suppose, for instance, that all background rates are approximately the same. Then, it is not reasonable to choose one pair over the others, since all of them are almost equally likely to be involved. Unfortunately, this is a general property of model (9). It will always either assign all of the weight to one pair for each incomplete event, never creating any distributions, or will give a degenerate solution. Indeed, the normalization constraints and the objective function, in each of its arguments, are all linear.

Model (8) does not have such a drawback for this example. It does not just pick the most likely participants of the event, but assigns weights to all pairs indicating how likely each of them is to be involved. This can be thought of as some sort of regularization property.

5.3. Discussion

As we mentioned in section 4, the objective function in (8) can be thought of as a sum of the energies of the events. Formally, if we ignore constant terms, it consists of two parts: quadratic terms, corresponding to the interaction of the incomplete events, and linear terms, corresponding to the energy of the incomplete events in the presence of the complete events and background rate values. The examples above were targeted to examine these parts separately to reveal their roles in the reconstruction process.

In the first example, we considered the quadratic part of the energy. We have seen that the incomplete events tend to gather on those timelines where their interaction energy is highest, which leads to aggressive cluster formation up to assigning all the weights to the same pair of agents.

On the other hand, the linear terms express the influence of the complete events and background rates, and do not allow the incomplete events to deviate too much from already existing clustering structure. Moreover, they regularize the solution, which represents the degree of uncertainty in the prediction, as demonstrated in the second example.

The methods arising from l^1-normalization (9) and from the alternative energy functional (7) have each shown some undesirable properties in these examples, and we will not consider them further. Of course, one is not restricted to only using l^1 or l^2 normalization, and one could consider a general l^p normalization of the weights or look at a hybrid constraint consisting of both l^1 and l^2 terms (or l^p terms). In the hybrid case, a constraint of the form

$$\sum_{\alpha\beta} f\left(m_i^{\alpha\beta}\right)^2 + (1-f)m_i^{\alpha\beta} = 1$$

$$\forall i = 1, \ldots, n,$$

could be employed, where $f \leqslant 1$ would represent how much emphasis to put on the l^2 term or the l^1 term. Although, for simplicity, we do not consider such a constraint in this work, it remains a potential avenue for future exploration.

6. Analysis

Note from figure 2 that the white circles naturally form a $K \times n$ matrix and our goal is to determine its entries. We denote the matrix as $\mathbf{X} = \{x_{ij}\}$. For future reference, it will be useful to express $\mathbf{X}$ in terms of its rows and columns

$$\mathbf{X} = \begin{pmatrix} \mathbf{r}_1^T \\ \vdots \\ \mathbf{r}_K^T \end{pmatrix} = \begin{pmatrix} \mathbf{c}_1 & \cdots & \mathbf{c}_n \end{pmatrix}. \tag{22}$$

Using these notations, problem (8) can be written as

$$\begin{cases} \sum_{i=1}^K \mathbf{r}_i^T \mathbf{A}_i \mathbf{r}_i + \mathbf{r}_i^T \mathbf{b}_i \to \max \\ \mathbf{c}_j^T \mathbf{c}_j = 1, \quad \forall\, j = 1, \ldots, n \\ x_{ij} \geqslant 0, \quad \forall\, i = 1, \ldots, K, \quad \forall\, j = 1, \ldots, n. \end{cases} \tag{23}$$

Here, $\mathbf{A}_i = \left\{a_i^{jl}\right\}$ is the symmetric $n \times n$ matrix of the interaction coefficients between the incomplete events on the ith timeline, and $\mathbf{b}_i = \left\{b_i^j\right\}$ is the column of size n of the energy coefficients for the incomplete events in the presence of the complete events and background rate on the ith timeline. Clearly, the entries of $\mathbf{A}_i$ and $\mathbf{b}_i$ are non-negative, for all $i = 1, \ldots, K$.

Theorem. *For problem (23):*

(i) There exists a global maximizer.
(ii) Every local maximizer (or even a stationary point) is a global maximizer.
(iii) If all b_i^j are strictly positive, then the maximizer is unique.

Proof. The objective function is continuous and the admissible set, given by the constraints, is compact. This proves (i). Define $y_{ij} = x_{ij}^2$. Then, problem (23) becomes

$$\begin{cases} \sum_{i=1}^K \left[\sum_{j,l=1}^n a_i^{jl} \sqrt{y_{ij} y_{il}} + \sum_{j=1}^n b_i^j \sqrt{y_{ij}} \right] \to \max \\ \sum_{i=1}^K y_{ij} = 1, \quad \forall\, j = 1, \ldots, n \\ y_{ij} \geqslant 0, \quad \forall\, i = 1, \ldots, K, \quad \forall\, j = 1, \ldots, n \end{cases}. \tag{24}$$

The admissible set in (24), given by the constraints, is convex. We will show that the objective function is concave on it, and strictly concave if all b_i^j are strictly positive, which implies (ii) and (iii).

Note that $a_i^{jl} \sqrt{y_{ij} y_{il}}$ is concave for all $i = 1, \ldots, K$ and $j, l = 1, \ldots, n$. This follows from the fact that for all $a, b, c, d \geqslant 0$ and $0 < \lambda < 1$ we have

$$\sqrt{(\lambda a + (1-\lambda)c)(\lambda b + (1-\lambda)d)} \geqslant \lambda \sqrt{ab} + (1-\lambda)\sqrt{cd}. \tag{25}$$

Indeed, squaring both sides of (25) gives a Cauchy-type inequality

$$cb + ad \geqslant 2\sqrt{abcd}, \tag{26}$$

after simplification. Now it suffices to show that the function

$$f_j(y_{1j}, \ldots, y_{Kj}) = \sum_{i=1}^K b_i^j \sqrt{y_{ij}} \tag{27}$$

is concave for all $j = 1, \ldots, n$. That is,

$$\sum_{i=1}^{K} b_i^j \sqrt{\lambda \hat{y}_{ij} + (1-\lambda)\check{y}_{ij}} \geqslant \sum_{i=1}^{K} b_i^j [\lambda \sqrt{\hat{y}_{ij}} + (1-\lambda)\sqrt{\check{y}_{ij}}] \tag{28}$$

for all admissible distinct $\{\hat{y}_{ij}\}_{i=1}^K$, $\{\check{y}_{ij}\}_{i=1}^K$ and $0 < \lambda < 1$. We further wish to show that (27) is strictly concave, that is inequality (28) must be strict, if all b_i^j are strictly positive. But both are true since the function $\sqrt{x}$ is strictly concave on $\{x : x \geqslant 0\}$. This completes the proof. □

If all pairs are active, then all background rates are nonzero, and we automatically have all b_i^j strictly positive, which implies the uniqueness of prediction in accordance with the theorem. When some pairs are inactive, part (iii) of the theorem is not applicable directly. Indeed, if for example timeline i is inactive, then there are no complete events on it and the corresponding background rate is 0; hence, $\mathbf{b}_i = \mathbf{0}$. Note however that in this case adding the constraint $\mathbf{r}_i = \mathbf{0}$, or simply excluding the timeline i from consideration, gives a problem with a smaller unknown matrix equivalent to (23). Thus, if we eliminate all inactive pairs in this way, we get a problem with all pairs in question being active, which guarantees the uniqueness of prediction.

So far, we implicitly assumed that we had no information at all about the participants of the incomplete events and each pair was considered as a possible candidate for prediction. Of course, if one of the participants of an event is known, then the pairs without this agent cannot be in involved, and the corresponding entries of $\mathbf{X}$ must be equal to 0, which means that we have additional constraints of the form $x_{ij} = 0$ for problem (23). These constraints however do not affect the convexity of the admissible set in the coordinates $y_{ij} = x_{ij}^2$. Therefore, all results of the theorem remain valid.

7. Results

In this section, we present and discuss the results of various tests of the proposed reconstruction method. Since the data from the Los Angeles gang network are incomplete, and the ground truth and interaction parameters for it are unavailable, it is not quite suitable for this purpose. Instead, we generate synthetic data using a Hawkes process (3), throw out some of the data at random, and then apply our algorithm to reconstruct it.

To evaluate the performance of our algorithm, we only focus on the *ordering* of various $m^{\alpha\beta}$ for each incomplete event i. Specifically, we determine for each incomplete event i the weights m_i for that event on various timelines, order them from lowest to highest, and find the corresponding rank of the ground truth timeline for that event. This is performed for two major reasons. First, our method (8) does not assign proper probabilities to various timelines, only weights that should be interpreted as being related to probability in a monotonic way. Second, from an operational point of view, the authorities are not very concerned with the actual probabilities with which each gang committed a given crime, but rather with a simple ranking of gangs from most likely to least likely, to prioritize their investigation.

As a first step, we compare our continuous method (8) to two others: one derived from the likelihood function (4) and one using the discrete model (10). However, note that methods (4) and (10) provide likelihoods (or energies) only for full allocations $\mathcal{A}$ of incomplete events, rather than one likelihood for each timeline per event. To bypass this issue, we simply define the likelihood $\hat{m}_i^{\alpha\beta}(f)$ that incomplete event i belongs to timeline $\alpha\beta$ under metric f to be

$$\hat{m}_i^{\alpha\beta}(f) = \sum_{\mathcal{A}_i^{\alpha\beta}} f(\mathcal{A}), \tag{29}$$

Table 1. Continuous method (8) compared to methods (4) and (10), for $N = 40$, $n = 4$, $k = 4$, $K = 6$, $\mu = 10^{-2}$ days^{-1}, $\omega = 10^{-1}$ days^{-1} and $\theta = 0.5$.

Method	Top 1	Top 2	Top 3	Top 4	Top 5
(4)	47.3%	68.1%	79.8%	87.7%	94.0%
(8)	47.1%	68.1%	79.7%	87.6%	94.1%
(10)	47.0%	68.1%	79.7%	87.6%	94.0%

where $\mathcal{A}_i^{\alpha\beta}$ is meant to represent only those allocations in which incomplete event i is attributed to timeline $\alpha\beta$, and $f = \mathcal{L}$ for (4) and $f = \Lambda$ for (10).

As mentioned previously, methods (4) and (10) are of combinatoric complexity, so we limit our testing here to a relatively small system with $N = 40$, $n = 4$, $k = 4$, $K = 6$. Here, we assume no knowledge of the participants in incomplete events, so each may be assigned to any of the $K = 6$ timelines. Simulations were run 10 000 times using parameters $\mu = 10^{-2}$ days^{-1}, $\omega = 10^{-1}$ days^{-1} and $\theta = 0.5$ for each pair of agents, which have approximately the order of magnitude estimated in [9]. Each simulation generated a ranking of the timelines for each incomplete event, and the percentages of incomplete events whose ground truth timelines were given certain ranks are shown in table 1.

Note that the three methods perform almost identically, each placing the correct timeline at top likelihood approximately 47% of the time, in the top-two likelihoods approximately 68% of the time and in the top-three likelihoods approximately 80% of the time. Since method (8) yields nearly indistinguishable solutions to those of (4) and (10), but is vastly more computationally effective, we focus only on this continuous method for the remainder of this section.

We next test our continuous method using datasets that more closely mimic the gang rivalry data. In all the experiments below, we have exactly one participant unknown for every incomplete event, which is the case for most of the gang data. Also, unless specified otherwise, we assume full connectedness of the network graph and use the same interaction parameters for each pair of agents as used above: $\mu = 10^{-2}$ days^{-1}, $\omega = 10^{-1}$ days^{-1} and $\theta = 0.5$.

Table 2 demonstrates the performance of the continuous method (8). It is organized as follows. The first three columns describe the dimensions of the network and the data the method was applied to, and the last three indicate how often, on average, a ground-truth unknown pair was in the top-one, top-two and top-three weights of the predicted distribution. The $\star$ value of k corresponds to the real Los Angeles gang network (see figure 3), which is not a fully connected graph. The 'Guessing' rows show the results that would be obtained by random guessing.

First we note that, in terms of prediction quality, the Los Angeles gang network roughly corresponds to a fully connected 6-nodes graph. This actually makes sense, since each gang has about five rivalries on average. Second, the prediction results depend rather mildly on the fraction of incomplete events, which implicitly confirms the fact that reconstruction model (8) captures the qualitative features of interaction process (3) rather well.

As for the results themselves, we can see that they are significantly better than those obtained by just random guessing. At the same time, they are not perfect. To see why this is so, we need to have a closer look at how they depend on the parameters of the system: μ, ω and θ. If self-excitation is too weak, that is, $\omega/\mu \ll 1$ and $\theta \ll 1$, then rate (3) will always remain near μ and the clusters will be vague and widespread. Hence, the method will give almost uniform distributions of weights, and choosing the pair with the biggest weight will be equivalent to random guessing. On the other hand, if self-excitation is very strong, that is,

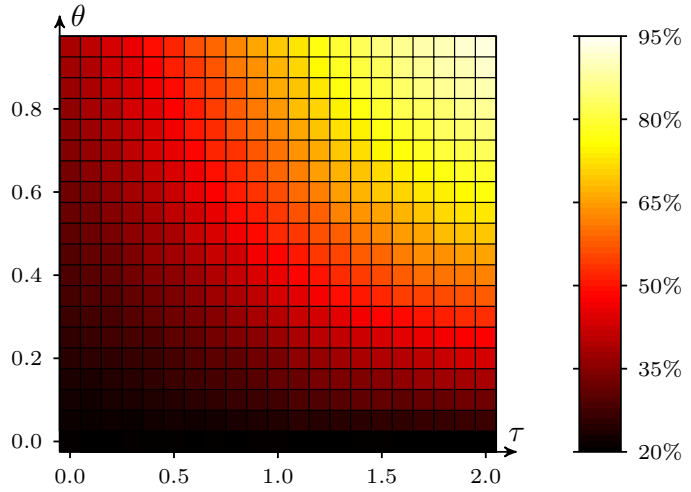

Figure 7. Dependence of the average percentage of correct predictions, obtained by choosing the pair with the highest weight for each distribution vector, on θ and $\tau = \log_{10}(\omega/\mu)$, for a fully connected six-agent network, with $N = 400$ and $n = 100$.

Table 2. Continuous model (8) performance results. The first three columns describe the dimensions of the network and the data the method was applied to, and the last three indicate how often, on average, a ground-truth unknown pair was in the top-one, top-two and top-three weights of the predicted distribution. The $\star$ value of k corresponds to the real Los Angeles gang network; see figure 3, which is not a fully connected graph. The 'Guessing' rows show the results that would be obtained by random guessing.

k	N	n	Top 1	Top 2	Top 3
5	400	50	57%	80%	92%
5	400	100	56%	79%	91%
5	400	200	54%	76%	90%
5	Guessing		25%	50%	75%
7	400	50	47%	69%	82%
7	400	100	46%	68%	80%
7	400	200	45%	65%	77%
7	Guessing		17%	33%	50%
9	400	50	42%	62%	73%
9	400	100	41%	60%	72%
9	400	200	39%	57%	69%
9	Guessing		13%	25%	38%
$\star$	400	50	50%	72%	83%
$\star$	400	100	49%	71%	82%
$\star$	400	200	48%	68%	80%

$\omega/\mu \gg 1$ and $\theta \simeq 1$, then the clusters will be sharp, the distribution vectors will be sparse, and choosing the pair with the biggest weight will give a reliable prediction.

Figure 7 confirms the above reasoning. Here, we applied our method to a fully connected six-agent network, with $N = 400$ and $n = 100$, varying the values of θ and $\tau = \log_{10}(\omega/\mu)$. For each distribution vector of weights, we simply picked the timeline with the highest weight and plotted the average percentage of correct predictions obtained in this way.

8. Conclusion

Retaliatory gang violence is a major problem in many metropolitan areas around the globe, and to curtail such violence, law enforcement agencies need to know who the participants were in a given altercation. We have shown that, under the assumptions that retaliatory violence on a gang network follows a Hawkes process of form (3), incomplete data on the participants of the offenses can be reconstructed using a computationally effective algorithm that maximizes an energy functional under a set of constraints—method (8). Moreover, when focusing on the likelihood rankings of gangs for incomplete events, method (8) seems to perform on par with a more probability-based algorithm (4) that is too complex to use on realistically sized datasets. Finally, we have shown that the performance of our method is deeply connected to the parameters of the Hawkes process in question, and in certain regimes may predict the correct participants with very high likelihood.

Of course, there are issues to overcome if our method is to be used on actual gang violence data, rather than on simulated events. First, for real datasets, the parameters of the process must be estimated from the events, rather than being known *a priori*. One could imagine accomplishing this in an iterative way: use the complete events to estimate parameters, use these parameters to estimate participants in unknown events, and then use these estimates to re-estimate the parameters, continuing the cycle until convergence (if convergence is indeed obtained). To implement this, however, one must choose how to use the estimated participants of events when re-estimating the interaction parameters, something that is not entirely clear given that our estimates of the participants are not probabilities. This could perhaps be accomplished via the expectation–maximization algorithm [25]. In this case, one would first have to consider the real likelihood function, incorporating this time both missing events and unknown parameters, and then come up with a suitable approximation for it that would make the problem less computationally expensive and yield plausible results.

Second, in real datasets one must be concerned with systematic deviation between the data and actual occurrences. Certain types of gang violence may be chronically under-reported in ways that will skew the detection of self-excitation or cause events to be allocated in an improper way. A thorough understanding of how this might affect our estimates should be had before trusting the results completely.

Acknowledgments

This work was supported by NSF grant DMS-0968309, ARO grant 58344-MA, ONR grant N000141010221 and AFOSR MURI grant FA9550-10-1-0569. The authors would like to thank Jeffrey Brantingham for providing comments on the later drafts of this paper, Captain Sean Malinowski of the Los Angeles Police Department for helpful conversations, and George Tita and the LAPD for providing the field data that motivated this work.

References

[1] Schafer J L and Graham J W 2002 Missing data: our view of the state of the art *Psychol. Methods* **7** 147–77
[2] Kossinets G 2006 Effects of missing data in social networks *Soc. Netw.* **28** 247–68
[3] Huisman M 2009 Imputation of missing network data: some simple procedures *J. Soc. Struct.* **10** (http://www.cmu.edu/joss/content/articles/volindex.html)
[4] Hoff P D, Raftery A E and Handcock M S 2002 Latent space approaches to social network analysis *J. Am. Stat. Assoc.* **97** 1090–8
[5] Hoff P D 2009 Multiplicative latent factor models for description and prediction of social networks *Comput. Math. Organ. Theory* **15** 261–72

[6] Huisman M and Steglich C E G 2008 Treatment of non-response in longitudinal network studies *Soc. Netw.* **30** 297–308
[7] Burt R S 1987 A note on missing network data in the general social survey *Soc. Netw.* **9** 63–73
[8] Tita G 2003 *Reducing Gun Violence: Results from an Intervention in East Los Angeles* (Santa Monica, CA: RAND Corporation)
[9] Egesdal M, Fathauer C, Louie K and Neuman J 2010 Statistical modeling of gang violence in Los Angeles *SUIRO* **3** 72–94 (available at http://www.SIAM.org/students/siuro/vol3/S01045.pdf)
[10] Hawkes A G 1971 Spectra of some self-exciting and mutually exciting point processes *Biometrika* **58** 83–90
[11] Hawkes A G and Oakes D 1974 A cluster process representation of a self-exciting process *J. Appl. Probab.* **11** 493–503
[12] Ogata Y 1988 Space-time point process models for earthquake occurrences *Ann. Inst. Stat. Math.* **50** 379–402
[13] Zhuang J, Ogata Y and Vere-Jones D 2002 Stochastic declustering of space-time earthquake occurrences *J. Am. Stat. Assoc.* **97** 369–80
[14] Tita G and Ridgeway G 2007 The impact of gang formation on local patterns of crime *J. Res. Crime Delinquency* **44** 208–37
[15] Jacobs B A and Wright R 2006 *Street Justice: Retaliation in the Criminal Underworld* (Cambridge: Cambridge University Press)
[16] Short M B, D'Orsogna M R, Brantingham P J and Tita G E 2009 Measuring and modeling repeat and near-repeat burglary effects *J. Quant. Criminol.* **25** 325–39
[17] Mohler G O, Short M B, Brantingham P J, Schoenberg F P and Tita G E 2011 Self-exciting point process modeling of crime *J. Am. Stat. Assoc.* **106** 100–8
[18] Lewis E, Mohler G, Brantingham P J and Bertozzi A 2010 Self-exciting point process models of insurgency in Iraq *UCLA CAM Reports 10-38* (doi:10.1057/sj.2011.21)
[19] Short M, Mohler G, Brantingham P J and Tita G 2010 Gang rivalry dynamics via coupled point process networks
[20] Short M, D'Orsogna M, Pasour V, Tita G, Brantingham P, Bertozzi A and Chayes L 2008 A statistical model of criminal behavior *Math. Models Methods Appl. Sci.* **18** 1249–67
[21] Rutenbar R A 1989 Simulated annealing algorithms: an overview *IEEE Circuits Devices Mag.* **5** 19–26
[22] B A Berg 2004 *Markov Chain Monte Carlo Simulations and Their Statistical Analysis* (Singapore: World Scientific)
[23] Ogata Y 1978 The asymptotic behaviour of maximum likelihood estimators for stationary point processes *Ann. Inst. Stat. Math.* **30** 243–61
[24] Ogata Y and Tanemura M 1984 Likelihood analysis of spatial point patterns *J. R. Stat. Soc.* B **46** 496–518 (available at http://www.jstor.org/stable/2345690)
[25] Dempster A P, Laird N M and Rubin D B 1977 Maximum likelihood from incomplete data via the EM algorithm *J. R. Stat. Soc.* B **39** 1–38 (available at http://www.jstor.org/pss/2984875)