Content-based Modeling of Reciprocal Relationships using Hawkes and Gaussian Processes

Xi Tan [1], **Syed A. Z. Naqvi** [1], **Alan (Yuan) Qi** [1,2], **Katherine A. Heller** [3], and **Vinayak Rao** [2]

[1]Department of Computer Science, Purdue University, West Lafayette, IN 47907
[2]Department of Statistics, Purdue University, West Lafayette, IN 47907
[3]Department of Statistical Science, Duke University, Durham, NC 27708

Abstract

There has been growing interest in inferring implicit social structures using interaction data. This approach is motivated by the fact that entities organize themselves into groups having frequent interactions between each other. Unlike previous approaches that focused on subjectively declared relationships, the idea is to exploit the actual evidence at hand to reach conclusions about group formations, resulting in more objective data-driven inferences. To this end, [5] have employed Hawkes processes, and proposed a Hawkes IRM model to infer social structures from interaction data. A major factor that encourages the use of Hawkes processes is the capability to model reciprocity in the interaction between social entities. However, reciprocation is dynamically conditioned upon two key factors: the significance of each message sent by the sender, and the receptivity to each message received by the receiver. In the model proposed by [5], reciprocity is not affected by either of these factors, since the content of each message is not taken into account. In this paper, we extend the work of [5] by introducing Gaussian processes (GPs) into the Hawkes IRM model: based on the content of each message, GPs are used to model the message significance as well as receptivity. This allows us to more accurately capture the interactions among entities. The application of GPs also allows us to flexibly model the rates of reciprocal activities between two entities, allowing asymmetry in reciprocity to be captured more accurately. This leads to better cluster detection capability. Our model outperforms previous Hawkes and Poisson process-based models at predicting verbal, email, and citation activities.

1 INTRODUCTION

In the social sciences, group dynamics is the study of the content and dynamics of the complex interactions occurring within a social group or between social groups. The study of group dynamics helps understand decision making processes, disease epidemics and develop effective therapeutic/control techniques. Early approaches [12, 19, 4] have focused on declared relationships between individuals to infer latent group structures. For example, if three people declare they like each other but dislike others, it is reasonable to put them into one group. However, these declared relationships are not easily accessible, sometimes incorrect and usually highly subjective. Another limitation of previous models is their incapability to capture reciprocity in social interactions. Reciprocity is a common characteristic in group dynamics. It expresses the fact that social entities reciprocate in their interaction between each other. For example, if Alice has sent a message to Bob, it increases the likelihood of Bob replying back to Alice. Reciprocity is expected to be more prominent between entities within a group, and hence it can be used to infer social groups.

To address these issues, recently, there has been a trend to infer implicit social structures using interaction data. This approach is motivated by the fact that interactions between different groups varies in nature and frequency. Unlike approaches that focused on subjectively declared relationships, the idea is to exploit the actual evidence at hand to reach conclusions about group formations, making this approach is more objective in nature. Recently, [5] proposed a nonparametric Bayesian model that is built upon mutually-exciting point processes, known as Hawkes processes [9, 10], and the Infinite Relational Model (IRM) [19, 4] to infer social structures from continuous time interaction data. Pairs of mutually-exciting Hawkes processes are able to exploit reciprocity to infer social groups; here the processes excite one another through their actualized events.

However, reciprocation is dynamically conditioned upon two key factors: the significance of each message sent by

the sender, and the receptiveness of the receiver to each incoming message. In real communication, conveying an important message develops interest in the receiver. Then, if the receiver finds the message relevant, reciprocation takes place. Accordingly, reciprocal communication emerges from the interplay of these two factors. The model proposed by [5] does not take these factors into consideration, instead assuming that entities reciprocate simply because they received a message, and giving no consideration to the content of the message and its effects on the interaction.

In this paper, we extend the work of [5] by introducing Gaussian processes (GPs) into the Hawkes IRM model. We use these to account for the content of the messages, capturing the message significance as well as receptivity. This allows us to more accurately capture the interactions among entities. The interaction between a pair of clusters is modeled as the additive effect of the interactions between all pairs of nodes in the two clusters, allowing us to identify the contribution of each pair of nodes, where the actual communication is taking place, to the interaction between a pair of clusters. The introduction of GPs also allows us to flexibly model the rates of reciprocal activities between two entities, hence the asymmetry in reciprocity can be captured more accurately. We show how this leads to a better cluster detection capability. Since our proposed work is a natural extension of Hawkes IRM, it covers both Poisson processes and IRM as special cases.

The remainder of the paper is organized as follows: section 2 discusses Poisson and Hawkes processes, with and without IRM. Section 3 describes our extension of the Hawkes IRM model. Section 4 presents an inference algorithm for our model, section 5 discusses related work, and section 6 presents experimental results using our model on synthetic, verbal, email, and citation data.

2 BACKGROUND

We start with a brief description of Poisson processes, Hawkes processes, and Hawkes IRM model.

2.1 Poisson and Hawkes Processes

Point processes are stochastic processes, realizations of which are collections of points in time or space. The former are called temporal point processes, and the latter, spatial point processes. The homogeneous Poisson process is the simplest example of a point process, have a constant rate function, while the inhomogeneous Poisson process has rate function λ varying with, say, time. Both are examples of completely random measures, where events in disjoint sets are independent of each other. Hawkes processes, on the other hand, are mutually-exciting doubly point processes, whose rate function is itself a stochastic process, depending on events of its own and of other processes.

For both Poisson processes and Hawkes processes, with conditional rate function $\lambda(t)$ and event time history $\mathcal{H}_{(0,T]} = \{t_1, \cdots, t_n\}$, the likelihood function can be written as

$$\mathcal{L}(\lambda(t)|\mathcal{H}) = \exp\{-\Lambda(0,T)\} \prod_{i=1}^{n} \lambda(t_i) \tag{1}$$

where $\Lambda(0,T) = \int_0^T \lambda(t)dt$ is the cumulative conditional rate function. When the conditional rate function $\lambda(t) = \lambda$ is a constant, the Poisson process likelihood is simply:

$$\mathcal{L}(\lambda|\mathcal{H}) = \exp\{-\lambda T\}\, \lambda^n \tag{2}$$

For a Hawkes process, the rate function λ depends on earlier events. Let $N(\cdot)$ and $N'(\cdot)$ be a pair of mutually-exciting Hawkes processes. The conditional rate function $\lambda(t)$ of $N(\cdot)$, given the event time history $\mathcal{H}_{N'} = \{t'_1, \cdots, t'_n\}$ of N', has the form

$$\lambda(t) = \gamma + \int_{-\infty}^{t} g(t-s)dN'(s) \tag{3}$$

where γ is the base rate of $N(\cdot)$, and the triggering function $g(\cdot)$ is a non-negative function such that $\int_0^\infty g(s)ds < 1$, ensuring that $N(\cdot)$ is stationary.

If $g(\cdot) = 0$ then the process becomes a Poisson process with rate γ. If the counting measure $N'(\cdot)$ is $N(\cdot)$ itself, the process is self- exciting: its current rate only depends on its own past events. If the two counting measures are different, the rate is affected by the past events of each other.

2.2 Hawkes Processes with Infinite Relational Model (HP+IRM)

Amongst the models that use declared relationships to infer group information, the Infinite Relational Model (IRM) [12] is especially flexible and popular. [5] has combined the IRM idea with the concept of Hawkes Processes to model reciprocity in the interaction between entity groups. Let V denote the vertices of the graph, corresponding to individuals. Then the generative model for a Hawkes process is defined as follows:

$$\pi|\alpha \sim CRP(\alpha) \tag{4}$$

$$\lambda_{pq}(t)|\gamma_{pq}, \beta_{pq}, \tau_{pq} = \gamma_{pq} n_p n_q \quad + \int_{-\infty}^{t} g_{pq}(t-s)dN_{qp}(s) \;\; \forall p, q \in range(\pi) \tag{5}$$

$$N_{pq}(\cdot)|\lambda_{pq} \sim HawkesProcess(\lambda_{pq}) \tag{6}$$

$$N_{uv}(\cdot)|N_{\pi(u)\pi(v)}, \pi \sim Thin(N_{\pi(u)\pi(v)}) \;\; \forall u, v \in V \tag{7}$$

Here π is a partition of the vertices V , distributed according to the Chinese restaurant process (CRP) with concentration parameter α. We use p and q to index clusters

of π. We denote the cluster that vertex u belongs to as $\pi(u)$. The operator $Thin$ refers to thinning; this means distributing the atoms of $N_{pq}(\cdot)$ among each $N_{uv}(\cdot)$, such that $N_{pq} = \sum_{u,v} N_{u,v}(\cdot)$. Any thinning scheme may be used, such as a uniform thinning, which uniformly picks to elements of a cluster. The type of reciprocation (parameterized by g_{pq} and g_{qp}, respectively) differs with events from cluster p to cluster q and events from cluster q to cluster p. This difference in reciprocity is what the model exploits to learn about social groups.

3 HAWKES PROCESSES WITH IRM AND GAUSSIAN PROCESSES (HPGP + IRM)

We define the Hawkes process conditional rate function as:

$$\lambda_{uv}(t) = \gamma_{pq} + \int_0^t \beta_{uv} e^{-\frac{t-s}{\tau_{uv}}} dN_{vu}(s) \tag{8}$$

where $p = \pi^{-1}(u), q = \pi^{-1}(v)$ are the clusters individuals u and v belong to; and the triggering function $g_{uv}(\cdot)$ is defined as:

$$g_{uv}(\delta) = \beta_{uv} e^{-\frac{\delta}{\tau_{uv}}} \tag{9}$$

Geometrically, the excitation function β_{pq} is essentially the "jump size" of the rate function $\lambda_{uv}(t)$ whenever a new message is received. However, in the above definition, β_{uv} is not affected by the content of the message; its value does not change based on the significance and receptivity of the messages.

We would like to define β_{uv} in a way such that it measures the significance and receptivity of individual messages communicated between individuals u and v. The content measure x_{vu} can be suitably defined according to the application, for example, it can be a distribution of words, the length of the message, or the text entropy of the message, etc. The individual level excitation function $\beta_{uv}(x_{vu}(s)) = 0$ if no message was sent from v to u at time s, but can be otherwise any non-negative function.

We propose to use two sets of Gaussian Process (GP) priors to address sources of inhomogeneity of the excitation functions $\beta_{uv}(\cdot)$, one for the significance of the message and one for the receptivity of the message:

$$\beta_{uv}(s) = e^{r_u(x_{vu}(s)) + s_v(x_{vu}(s))} \tag{10}$$

where

$$r_u(\cdot) \sim \mathcal{GP}(0, k_r) \tag{11}$$

$$s_v(\cdot) \sim \mathcal{GP}(0, k_s) \tag{12}$$

k_r and k_s are radial basis function (RBF) kernels of the GPs. The exponential transformation is used to make sure that $\beta_{uv}(\cdot)$ is non-negative.

Larger values of r_u and s_v indicate that an important message has been sent by the sender, and receiver is receptive to the message, these result in larger values for β_{uv}. If either r_u or s_v is small, or both of them have smaller values, it leads to smaller values of β_{uv}. Application of GP functions also allows us to flexibly model the rates of reciprocal activities between two entities, allowing the asymmetry in reciprocity to be captured more accurately. This, as a by-product, leads to a better cluster detection capability.

The receptivity and significance functions r_u and s_v may have different behaviors and hence are designed to come from two different GPs. One subtle point is that although r_u and s_v seem exchangeable in the definition of β_{uv} and both use message content x_{vu} as input, they are evaluated from different perspectives: r_u evaluates x_{vu} from the receiver u's perspective, while s_v from the sender v's perspective. One alternative way is to model a single pair of GPs $s(\cdot)$ and $r(\cdot)$ for all users, instead of this per-user GP $s_u(\cdot)$ and $r_v(\cdot)$ framework. Experiments have shown that both the modeling schemes have good performances, however, we believe that the per-user GP setting can reveal more interesting user-specific details, and hence in the later sections, our results are based on the per-user GP framework.

The generative process of our model can be summarized as:

$$\pi|\alpha \sim CRP(\alpha) \tag{13}$$

$$\lambda_{uv}(t)|\gamma_{pq}, \beta_{uv}(\cdot), \tau_{uv} = \gamma_{pq} + \int_{-\infty}^t \beta_{uv}(\mathcal{X}_{vu}) e^{-\frac{t-s}{\tau_{uv}}} dN_{vu}(s) \tag{14}$$

$$N_{uv}(\cdot)|\lambda_{uv} \sim HawkesProcess(\lambda_{uv}) \tag{15}$$

where $\mathcal{X}_{vu} = \{x_{vu}(s)\}$ is the set of all messages sent from v to u, and the cluster level excitation function β_{pq} can be seen as an additive effect of β_{uv}:

$$\beta_{pq}(\mathcal{X}_{qp}) = \sum_{\pi(u)=p, \pi(v)=q} \beta_{uv}(x_{vu}(s)) \tag{16}$$

Now, the excitation function β_{pq} is no longer a constant, as in [5], but a function of the message content in the past events of the reciprocal process N_{qp}, taking into account both the significance and the receptivity of the messages. Our model is a generalization of the model described in [5], and if β_{uv} in equation 10 are constants, our model reduces to the model described in [5]. Therefore, all the basic features of the original model are inherited by our model. Also, in our modeling framework, the individual rate function λ_{uv} is affected by the group initial rate γ_{pq}, which, on the one hand, tends to put similarly behaving individuals into the same cluster; and on the other hand, if one member of a group is heavily influenced by a particular message, it is highly likely that other individuals in the same group will also be affected.

3.1 Stability Conditions of HPGP + IRM

For Hawkes processes with constant excitation functions β_{pq}, the sufficient condition of stationarity is $\beta_{pq}\tau_{pq} < 1$, derived from the condition $\int_0^\infty \beta(s)ds < 1$. By contrast, since our β_{pq} is a function of message contents, the expectation of $\lambda(t)$ cannot be time invariant. Therefore, the stationarity condition no longer holds. However, since β_{pq} is evaluated at finite locations (in the domain of message content x), we can define β_{pq}^{MAX} to be the maximum value of β_{pq} across all locations. For our model, we can still require that $\beta_{pq}^{MAX}\int_0^\infty e^{-\frac{u}{\tau_{pq}}}du < 1$. Since $\beta_{pq}^{MAX}\int_0^\infty e^{-\frac{u}{\tau_{pq}}}du = \beta_{pq}^{MAX}\tau_{pq}$, we just need to make sure that $\beta_{pq}^{MAX}\tau_{pq} < 1$.

4 HPGP + IRM INFERENCE

We perform posterior inference using Markov chain Monte Carlo method. In our model there is no conjugacy between prior and the likelihood, hence we can not marginalize out parameters and must sample all of them separately. To infer the partition of individuals π, the concentration parameter α, the parameters of each Hawkes process $\theta_{pq} = \{\gamma_{pq}, \tau_{pq}\}$, the training and test point projections of functions r_u and s_v, we use Algorithm 5 in [15] to draw samples from the posterior. We use elliptical slice sampling [14] for r_u and s_v, and standard slice sampling [16] for γ_{pq}, τ_{pq} and α. In case of τ_{pq} we set the upper bound of the slice sampler to $\frac{1}{\beta_{pq}^{MAX}}$, to ensure that $\beta_{pq}^{MAX}\tau_{pq} < 1$.

5 RELATED WORK

The interest of modeling relational data dates back to at least the work of [11], who introduced the Bayesian formulation of the stochastic block-model. This model was then extended by [12] to the Infinite Relational Model (IRM).

The IRM typically assumes that there is a fixed graph, describing the relationship between individuals, which is observed. This idea is used in many proposed works [12, 19]. Our model does not make this assumption, but learns the relationship among participants' interactions.

There have also been research works modeling relational events via latent classes [6]. They assume each event's sender, receiver, and action type are conditionally independent given the latent class for that event. This strong assumption greatly simplifies the model, but may not reflect real situations. Our model is not limited to any fixed number of action types.

Other works [17, 18, 7] are based on temporal Poisson-processes, where the rate of events on each edge is independent of every other edge. Although [18, 7] allow mutually exciting events to be modeled, they do not use content information to model dependencies between events. Our model uses Hawkes processes which are capable of dealing with interaction and reciprocal events, and also use message content information to capture the interactions more accurately. Our work is also closely related to [13]. They combine mutually exciting Hawkes process with random graph models by defining the excitation function, between a pair of nodes, as a product of a latent binary indicator variable, indicating the presence or absence of edge, and weight variable that determines the strength of interaction between the two nodes. However, unlike our model, their method does not use side information, such as information content, and simply relies on time interaction data to infer latent network structures. Lastly, our work extends the work of [5]. In their paper, the excitation function is not affected by the information content of the message. By introducing Gaussian processes, we are able to model non homogeneous excitation functions. In addition to that, since we use Gaussian processes to model the flexible rates of reciprocal activities between two entities, our model can capture the asymmetry in reciprocity more accurately. This, as a by-product, leads to a better cluster detection capability. The model in [8] does not have this leverage.

6 EXPERIMENTS

We compared our model (HPGP + IRM) to four methods: 1) Poisson Process Model (Poisson), 2) Hawkes Process Model (HP), 3) Poisson Processes with IRM (Poisson + IRM), and 4) Hawkes Processes with IRM (HP + IRM).

6.1 Synthetic Data Sets

We tested several synthetic data sets under various conditions to compare different model fittings to the rate functions, as well as their clustering behaviors.

A Simple Case Consists of Two Individuals. To generate synthetic data set, we need to set parameter values γ_{uv}, and τ_{uv}, as well as the functional form of $\beta_{uv}(\cdot)$ and message content measure x_{vu}. In figure 1, two mutually-exciting Hawkes processes are simulated during time interval $(0, 10]$, where $\gamma_{12} = \gamma_{21} = 0.1$, $\tau_{12} = \tau_{21} = 1$.

In part (a), case 1 used a constant message content $x_{12}(t_i) = x_{21}(t_i') = 1$ for all event times t_i and t_i', and a constant excitation function $\beta_{12}(x) = \beta_{21}(x) = x = 1$ for all messages. Since this synthetic data set has constant β values, it is essentially generated from a HP+IRM; we see that HP+IRM and our model, a generalization to HP+IRM, both perform well, and are better than other models, in terms of log-likelihood shown in table 1.

In part (b), case 2 used the same settings as part (a), except for the introduction of variable message content, where both $x_{12}(t_i)$ and $x_{21}(t_i')$ follow an exponential distribution $\exp(0.5)$, which can be thought of as different message entropy values at different event times t_i and t_i'. We see that

the jump sizes of both processes are no longer constant. This cannot be modeled by a constant β model, but can only be handled by models like ours, which allow variable β. The effectiveness of our model in this case can be seen from the comparison of the log-likelihoods in table 1.

In part (c), case 3 further introduced non-constant $\beta_{uv}(\cdot)$, with all other settings being the same as in case 2, but $\beta_{12}(t_i) = e^{2sin(x_{21}(t_i))+1.5log(x_{21}(t_i))}$ and $\beta_{21}(t'_i) = e^{0.1cos(x_{12}(t'_i))+0.2\sqrt{x_{12}(t'_i)}}$, where $r_1(x_{21}(t_i)) = 2sin(x_{21}(t_i))$, $r_2(x_{12}(t'_i)) = 0.1cos(x_{12}(t'_i))$, $s_1(x_{12}(t'_i)) = 0.2\sqrt{x_{12}(t'_i)}$, and $s_2(x_{21}(t_i)) = 1.5log(x_{21}(t_i))$. Again, the jump sizes for both processes are not constant, and also note that $\beta_{21}(x) > \beta_{12}(x), \forall x \in (0, 10)$. This suggests that process 2 is excited to respond to any messages received from process 1, while process 1 is reluctant to respond to messages sent from process 2. In this case, the difference in log-likelihoods of different models is pronounced even more.

Table 1: Log likelihood comparison for the three-case synthetic data set

	CASE 1	CASE 2	CASE 3
HPGP+IRM	-21.88	-13.41	-10.86
HP+IRM	-22.97	-35.53	-82.78
POISSON + IRM	-72.31	-89.73	-126.33
HP	-129.37	-238.94	-192.78
POISSON	-127.83	-182.76	-187.23

Next, we will discuss our modeling preferences based on the three-case example used in figure 1.

GP Against Simple Parametric Functions. In order to demonstrate the effectiveness of using GP in our model, we compared its performances with simple parametric functions. In table 2, we summarize the log likelihood for the three-case synthetic data set mentioned earlier in figure 1, using GP and simple polynomials (up to order 3). The results clearly show the superior performance of GP over polynomial functions. The coefficients of polynomials are estimated by sampling from the posterior.

Table 2: Log likelihood comparison between GP and simple parametric functions

	GP	CUBIC	QUAD	LINEAR
CASE 1	-21.88	-38.67	-38.88	-39.18
CASE 2	-13.41	-61.27	-78.17	-89.28
CASE 3	-10.86	-71.26	-72.13	-76.73

Estimate Kernel Width From Data. In our experiment, we used the RBF (radial basis function) kernel, which has the form:

$$k(\delta) = \exp\left(-\frac{\delta^2}{2\sigma^2}\right) \tag{17}$$

where δ is the distance between two data points, and σ the kernel width. The estimation of the kernel width is crucial in our modeling framework as it controls the complexity of the underlying receptivity and significance functions. We applied 3 different approaches to estimate σ: Bayesian, heuristic, and fixed. The Bayesian approach introduces a prior on σ and obtains an estimate using MCMC; the heuristic way, bearing in mind that sigma largely depends on the maximum distance among the training data, estimates σ directly from sample data distances; and the fixed approach manually assigns a fixed value to the kernel width. It is evident from table 3 that the Bayesian approach is the best choice for our model in terms of log likelihood.

Table 3: Log likelihood comparison for kernel estimation

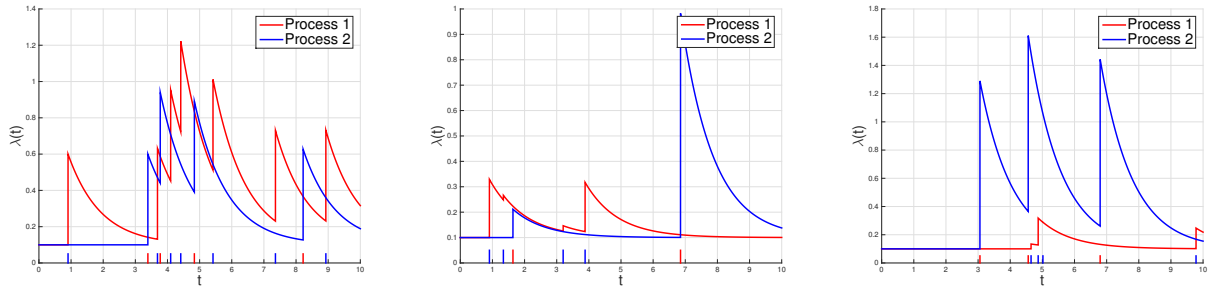
	BAYESIAN	HEURISTIC	FIXED
CASE 1	-21.88	-25.12	-39.78
CASE 2	-13.41	-17.16	-18.72
CASE 3	-10.86	-22.13	-24.67

Comparison Between Different Information Metrics. We compared four strategies to evaluate the information content of a message: KL divergence of word distribution, message length, TF-IDF, and message Shannon entropy. Using length as the measure of information may not be sufficient in practice; the importance of a message is simply determined by its longevity, without giving any consideration to the content. In case of Shannon entropy, however, the significance and receptivity of the message are better captured. TF-IDF has similar behavior and characteristics as those of message entropy. The best performance in our experiments were given by using KL divergence of word distribution and Shannon entropy, and we preferred KL divergence of word distribution over the other measures because it is more interpretable, and seemed to give consistent good performances in terms of log-likelihoods as shown in table 4. However, encoding content information efficiently is still an open question, and certainly a direction for future work.

Table 4: Log likelihood comparison for different information metrics

	WORD KL	ENTROPY	TF-IDF	LENGTH
CASE 1	-21.88	-21.98	-39.38	-128.76
CASE 2	-13.41	-12.78	-28.61	-87.21
CASE 3	-10.86	-12.63	-23.78	-72.13

Next, we will discuss a more detailed example consisting of three individuals.



(a) Case 1: x constant, β simple function $\beta = x$. The "jump sizes" are constant.

(b) Case 2: x random, β simple function $\beta = x$. The "jump sizes" are not constant.

(c) Case 3: x random, β non-trivial function. The "jump sizes" are not constant.

Figure 1: Simulated rate functions of two individuals

A Full Example Consists of Three Individuals. In this example, we put processes 1 and 2 in one cluster whereas process 3 is in another cluster, and we also intentionally made them behave differently to each other.

The settings we used were $m_{ij} \sim multinomial(p = [0.25, 0.25, 0.25, 0.25], n = 4), \forall i, j \in \{1, 2, 3\}$, which could represent a dialog consisting of only four words, and each m_{ij} can be thought of as the distribution of these four words in a message sent from j to i. We define the message content measure as $x_{ij} = KL(m_{ij}||\bar{m}_i)$, where $\bar{m}_i$ is the four-word distribution assigned to individual i ($\bar{m}_i = (1, 1, 1, 1), \forall i$ in our experiment). For the excitation functions we have: $\beta_{12} = \beta_{21} = 5\exp(1/x)$, $\beta_{23} = \beta_{31} = 0.1\exp(1/x)$, and $\beta_{13} = \beta_{32} = 10\exp(1/x)$. Note that $\beta_{12} = \beta_{21}$, $\beta_{31} < \beta_{13}$, and $\beta_{32} > \beta_{23}$.

Figure 2 (a) shows that processes 1 and 2 are frequently interacting in a similar way, while in part (b), process 3 is not excited to respond to messages from process 1 but tends to, suggested in part (c), reply to process 2's messages more actively. In figure 2 (g, h, and i), we see that only our model was able to correctly cluster processes 1 and 2 in the same cluster and put process 3 in a separate one. On the other hand, the other models generated redundant clusters. We have also shown in figure 2 (d, e, and f) that our model successfully recovered the underlying excitation functions.

6.2 Real Data Sets

We tested our model on various turn-taking data sets, which include public meetings, private conversations, email communications, and publication citations. Each data set has several lines of event records, indicated by a quadruplet (t_i, s_i, r_i, m_i), where t_i is the time when the event took place, s_i the index of the sender, r_i the index of the recipient, and m_i the message word distribution.

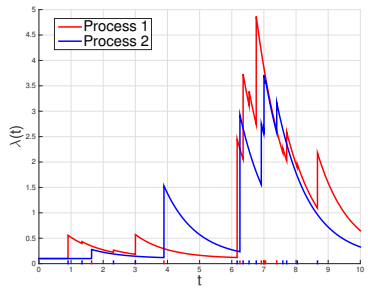
We divided the data set into two parts: the first part consists of the first 90% of the data lines, used as the training data set; and the second part contains the remaining 10% of the data lines, used as the testing data set. To compute the average log probability, we ran our code 10 times with different prior settings and computed the mean and standard deviation of the 10 values.

Enron email threads The Enron data set (ENRON) contains about half a million email messages sent or received by about 150 senior managers of the Enron corporation [2, 3]. We restricted ourselves to "true" conversation emails (e.g., auto-messages were ignored) sent and received only from the domain "@enron.com", and identified the threads by time, sender, receiver, and the subject line. The longest email communication was selected.

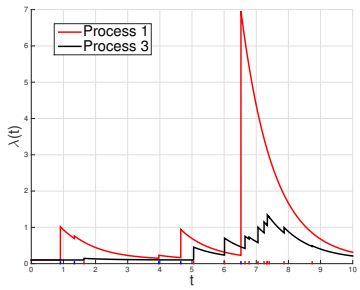
Santa Barbara Conversation Corpus The Santa Barbara Corpus [1] data set (SB) contains text and video recordings for various conversations. The data set used (#33) is a lively family argument/discussion recorded at a vacation home in Falmouth, Massachusetts. There are eight participants, all relatives or close friends. Discussion centers around a disagreement Jennifer (#2) is having with her mother Lisbeth (#5).

High-energy Physics Theory Citation Network The Arxiv HEP-TH (high energy physics theory) citation data set (CITATION) covers all 352807 citations of 27770 papers published during the time period January 1993 to April 2003 (124 months). We converted paper citation events to author citation events. For example, if a paper by authors A and B cited another paper by authors C, D, and E, we would record six events: A cited C, D, and E; and B cited C, D, and E. Only the most cited 17 authors and 97 citation events in the year 2003 were used from this data set.

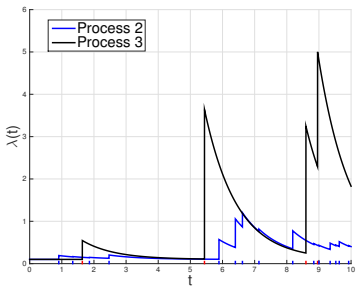
Results Table 5 and 6 show, for training and test data sets respectively, the predictive probability results as well as the most probable predictive number of clusters for competing methods. We used 10-fold cross-validation to prevent our model from being over-fitted to training data sets, and the performances on real data sets suggested a good generalization ability of our model.



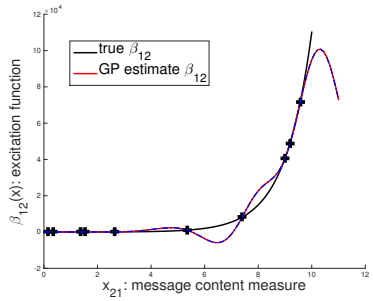
(a) $\beta_{12} = 5\exp(1/x)$
$\beta_{21} = 5\exp(1/x)$



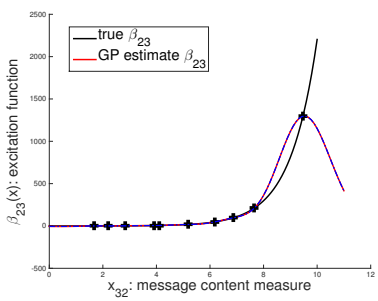
(b) $\beta_{13} = 10\exp(1/x)$
$\beta_{31} = 0.1\exp(1/x)$



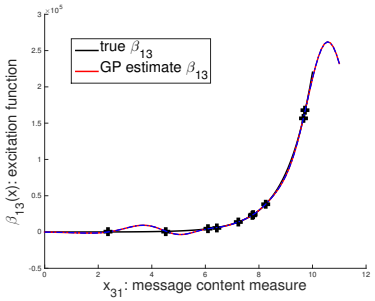
(c) $\beta_{23} = 0.1\exp(1/x)$
$\beta_{32} = 10\exp(1/x)$



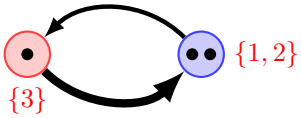
(d) GP plot of β_{12}



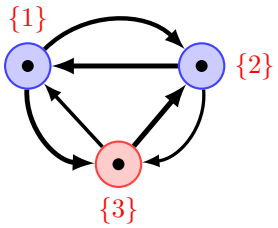
(e) GP plot of β_{23}



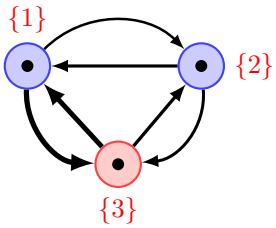
(f) GP plot of β_{13}



(g) HPGP+IRM



(h) HP + IRM



(i) Poisson + IRM

Figure 2: Simulated rate functions of three individuals and their cluster configurations

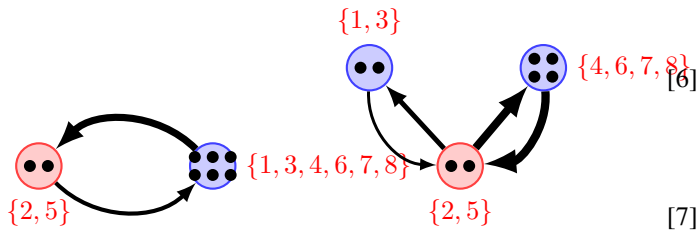
Table 5: Average log likelihood for each model with standard error (TRAINING data sets). N is number of individuals, T is number of events, and C the predicted number of clusters.

	ENRON	SB #33	CITATION
(N, T, C)	(2, 896, 2)	(8, 499, 8)	(17, 97, 17)
HPGP + IRM	5612.67 ± 0.13	672.03 ± 0.11	1265.31 ± 0.14
HP + IRM	5513.25 ± 0.12	475.13 ± 0.50	987.34 ± 0.23
POISSON + IRM	2360.37 ± 0.06	572.35 ± 0.11	918.56 ± 0.17

Table 6: Average log predictive likelihood for each model with standard error (TEST data sets).

	ENRON	SB #33	CITATION
C	2	2	11
HPGP + IRM	327.13 ± 0.02	126.87 ± 0.05	217.51 ± 0.43
HP + IRM	270.36 ± 0.01	89.05 ± 0.04	127.81 ± 0.32
POISSON + IRM	46.21 ± 0.01	13.08 ± 0.00	97.00 ± 0.41

We also compared our model with HP+IRM in terms of cluster detection capability. Figure 3 shows the cluster configurations generated by our model and HP+IRM. This dataset is a record of a lively family argument/discussion. There were eight participants, all relatives or close friends, but the main communication was between Jennifer (#2) and her mother Lisbeth (#5). For our model, Jennifer and Lisbeth were put in one cluster, and rest in the other. This is more consistent with data evidence: Jennifer and Lisbeth reciprocate each other more frequently, and respond occasionally to others, despite receiving a lot of messages from them. Individuals other than #2 and #5 may be further decomposed into subgroups, but at this level, the best clustering would probably be the one given by our model. The contrast in the thicknesses of the arrows between the two clusters correctly reveals this aspect. On the other hand, the cluster configuration generated by HP+IRM model indicates a high level of reciprocity, indicated by comparable thicknesses of the two arrows, between clusters $\{2,5\}$ and $\{4,6,7,8\}$ which is inconsistent with data evidence. Additionally, the model creates an extra cluster,$\{1,3\}$, which is inconsistent with data evidence.



(a) HPGP+IRM (b) HP+IRM

Figure 3: Diagram for data set *SB #33*. The thickness of the arrows are proportional to the expectation of the rate function.

7 CONCLUSION

In this paper, we have presented a non-parametric Bayesian model that uses Hawkes processes to model reciprocal relationships. Unlike previous approaches, our model utilizes the content of the messages to model reciprocity. Based on the content, our model captures the significance of the message sent by the sender, and receptivity to the message received by the receiver. This gives us the leverage to model reciprocity in a more realistic manner and more accurately. Empirical results suggest that our novel model formulation can yield improved predictive probability results, and can reveal clusters more accurately than alternative methods.

8 Acknowledgements

Xi Tan, Syed A. Z. Naqvi, and Dr. Alan (Yuan) Qi were supported by NSF CAREER award IIS-1054903, and the NSF Science and Technology Center.

References

[1] SB Data Set. `http://www.linguistics.ucsb.edu/research/santa-barbara-corpus`. [Online; accessed 20-August-2014].

[2] Enron Email Data Set (RDdata). `http://www.ahschulz.de/enron-email-data/enron_mysqldump.RData`, 2011. [Online; accessed 20-August-2014].

[3] Enron Email Data Set (Text). `https://www.cs.cmu.edu/~./enron/enron_mail_20110402.tgz`, 2011. [Online; accessed 20-August-2014].

[4] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.

[5] Charles Blundell, Jeff Beck, and Katherine A Heller. Modelling reciprocating relationships with hawkes processes. In *Advances in Neural Information Processing Systems*, pages 2600–2608, 2012.

[6] Christopher DuBois and Padhraic Smyth. Modeling relational events via latent classes. In *Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2010.

[7] Asela Gunawardanal, Christopher Meek, and Puyang Xu. A model for temporal dependencies in event streams. In *Neural Information Processing Systems (NIPS)*, 2011.

[8] Fangjian Guo, Charles Blundell, Hanna Wallach, and Katerine A. Heller. The bayesian echo chamber: Modeling influence in conversations. In *arXiv*, 2014.

[9] Alan G. Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 438–443, 1971.

[10] Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, pages 83–90, 1971.

[11] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic block models: first steps. In *Social Networks*, volume 5, pages 109–137, 1983.

[12] Charles Kemp, Joshua B Tenenbaum, Thomas L Griffiths, Takeshi Yamada, and Naonori Ueda. Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, page 5, 2006.

[13] Scott Linderman and Ryan Adams. Discovering latent network structure in point process data. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1413–1421, 2014.

[14] Iain Murray, Ryan P. Adams, and David Mackay. Elliptical slice sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 541–548, 2010.

[15] Radford M. Neal. Markov chain sampling methods for dirichlet process mixture models. Technical report, 1998.

[16] Radford M. Neal. Slice sampling. In *Annals of Statistics*, 2003.

[17] Uri Nodelaman, Christian R. Shelton, and Daphne Koller. Continuous time bayesian networks. In *Uncertainty in Artificial Intelligence (UAI)*, 2002.

[18] Shyamsundar Rajaram, Thore Grapple, and Herbrich Ralf. Poisson-networks: A model of structured point processes. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2005.

[19] Z Xu, V Tresp, K Yu, and HP Kriegel. Infinite hidden relational models. In *Proceedings of the 22nd International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2006.