Common Growth Patterns for Regional Social Networks: A Point Process Approach

Tiandong Wang[1,*] and Sidney I. Resnick[2]

[1]*Department of Statistics, Texas A&M University, College Station, TX 77843, U.S.A.*
[2]*School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853, U.S.A.*

Abstract

In this paper, we study macroscopic growth dynamics of social network link formation. Rather than focusing on one particular dataset, we find invariant behavior in regional social networks that are geographically concentrated. Empirical findings suggest that the startup phase of a regional network can be modeled by a self-exciting point process. After the startup phase ends, the growth of the links can be modeled by a non-homogeneous Poisson process with a constant rate across the day but varying rates from day to day, plus a nightly inactive period when local users are expected to be asleep. Conclusions are drawn based on analyzing four different datasets, three of which are regional and a non-regional one is included for contrast.

Keywords *estimation; network growth; non-homogeneous Poisson processes; point processes; self-exciting processes*

1 Introduction

Network growth can be measured with different methods, for instance by counting links or by counting nodes. A large number of links or edges can indicate that users find the network useful and compelling and may encourage potential users to join. A large number of nodes may be a driver of network revenue from subscriptions or advertising. Based on understanding growth patterns, network companies can decide how and when to conduct initiatives to maintain and enhance network attractiveness and usefulness. This paper focuses on statistically describing the process of edge growth applicable to certain regionally restricted social networks.

Social network analysis has triggered an exploding number of research topics. In the network and computer science literature, many novel models based on point processes have been proposed to analyze network evolution. For example, in Zhao et al. (2015), a self-exciting point process (SEPP) is used to predict the number of retweets for a given post on Twitter. Subsequently data-driven refinements of the SEPP model were proposed, e.g. Kobayashi and Tideh (2016); Rizoiu et al. (2017b); Mishra et al. (2016); Martin et al. (2016); Srijith et al. (2017). These studies emphasized user-based microscopic network evolution, e.g. the cascading of a given tweet and individual's social connectivity dynamics (Zang et al., 2017). For studying macroscopic evolution of networks over time, other approaches have been used to study densification and diameter growth. Leskovec et al. (2007) study several real networks and propose generative models that match the observed densification and diameter growth. Taking communities into consideration, Backstrom et al. (2006) have examined the growth pattern for a specific group of members in a network.

---

*Corresponding author. Email: twang@stat.tamu.edu.

One open question is how best to combine the observed growth patterns in real networks with existing standard network growth paradigms such as preferential attachment (Bollobás et al., 2003; Krapivsky et al., 2001; Krapivsky and Redner, 2001), the Bass model (Mahajan et al., 1990) and the susceptible-infected model (Anderson et al., 1992). These existing network models possess some theoretical tractability, but struggle to offer detailed information about link growth in real time especially when they model data sets that do not have complete timestamp records. Some resolutions have recently been proposed for example in Zang et al. (2016); Mislove et al. (2008) where the primary focus is on modeling the growth of a particular network, rather than looking for invariants across networks. Here we emphasize common properties across regional networks that are largely localized to one time zone, and focus on the growth of edges in the network, which is a measure of the utility network users find in a network. Using growth patterns observed in common across several data sets, we propose a realistic variant of the classical preferential attachment model; see Wang and Resnick (2020) for details.

Certain network data sets record link creation with the timestamp of when the link was created and it is natural to model network growth using these timestamps. In the graph representation of social networks, edges often represent individual interactions between two users. Some of these networks are regional in that users in the network are geographically concentrated (e.g. comments left on the talk page of articles in Dutch Wikipedia), compared to less geographically concentrated ones (e.g. comments left on the talk page of articles in English Wikipedia). Our point process approach captures some common patterns of network growth in regional social networks.

We regard times of edge creation as points or events in a point process. Initially network evolution can be modeled by a SEPP with an exponential triggering function but after this initial startup phase of the network ends, the point process is close to a non-homogeneous Poisson process (NHPP), apart from a nightly inactive period when most users in the same regional locale are asleep and hence not active on the network. We discuss four different network datasets listed on KONECT (konect.cc/networks):

- Facebook wall posts for users in New Orleans, available at http://konect.cc/networks/facebook-wosn-wall/.
- Dutch Wikipedia Talk, available at http://konect.cc/networks/wiki_talk_nl/.
- German Wikipedia Talk, available at http://konect.cc/networks/wiki_talk_de/.
- MathOverflow, available at http://konect.cc/networks/sx-mathoverflow/.

The first three networks are regional and instances of our findings but the last one is not and is included for contrast.

The rest of the paper is organized as follows. In Section 2, we summarize background knowledge on a SEPP and outline an estimation algorithm (cf. Kirchner, 2017; Kirchner and Bercher, 2018) which is used throughout the paper when estimating parameters in a SEPP. Analyses on the four network datasets are included in Section 3 and Section 4 highlights key findings in this study and raises open problems for future study.

2 Method

2.1 Specifying Self Exciting Point Processes (SEPPs)

An initial use of the SEPP was for modelling earthquake occurrences (Ogata, 1988), but subsequently the SEPP model found a variety of applications in finance (Bacry et al., 2015; Hawkes, 2018; Embrechts et al., 2011), social media (Zhao et al., 2015; Rizoiu et al., 2017a,b; Srijith

et al., 2017), terrorist activity (Porter and White, 2012), as well as crime and security (Mohler, 2013). Here is a brief overview of the SEPP.

A univariate point process is constructed on a probability space $(\Omega, \mathbf{P}, \mathcal{F})$ where points are viewed as events. Define on this probability space a sequence of stochastic event times $0 = T_0 \leqslant T_1 \leqslant T_2 \leqslant \cdots \leqslant T_n \leqslant T_{n+1} \leqslant \ldots$, and using the standard point measure notation

$$\epsilon_x(A) := \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{if } x \notin A. \end{cases}$$

the counting process on the Borel sets of $\mathbb{R}_+$, $\mathcal{B}(\mathbb{R}_+)$, is defined as

$$N(\cdot) := \sum_{i=1}^{\infty} \varepsilon_{T_i}(\cdot).$$

We denote $\mathcal{H}_t$ as the *history* associated with N, and use the notation $N(t) \equiv N([0, t])$, $t \geqslant 0$, in the rest of this paper.

A *self-exciting* point process (SEPP) or *Hawkes process* is a point process $N(\cdot)$ which has a conditional intensity (cf. Daley and Vere-Jones (2003)) consisting of

- A *baseline or background intensity* $\eta \geqslant 0$ and
- A *triggering function* $g : \mathbb{R}_+ \mapsto \mathbb{R}_+$ measuring how much a recent event increases the intensity.

We assume the conditional intensity has the following form:

$$\lambda(t|\mathcal{H}_t) = \eta + K \int_0^t g(t - s)\mathrm{d}N(s), \qquad \eta, K > 0, t \geqslant 0. \tag{1}$$

For brevity we often suppress writing the history $\mathcal{H}_t$ and simply use $\lambda(t)$ to denote the conditional intensity.

There are two common choices of the triggering function g, namely

- Exponential function:

$$g(t; \beta) = \beta e^{-\beta t}, \qquad \text{for some } \beta > 0, t \geqslant 0; \tag{2}$$

- Power-law function:

$$g(t; \gamma) = \frac{\gamma - 1}{(1 + t)^{\gamma}}, \qquad \text{for some } \gamma > 1, t \geqslant 0; \tag{3}$$

and we only focus on these two choices.

In Section 2.2 we outline an estimation procedure for the SEPP parameters based on Kirchner and Bercher (2018), which approximates a SEPP by an integer-valued autoregressive model with infinite order, INAR(∞). Theoretical justifications of the estimation procedure are given in Kirchner (2016), where the stationarity of both the SEPP and INAR(∞) is assumed. In fact, the stationarity of the SEPP is obtained from a branching representation as given in Hawkes and Oakes (1974) and with the triggering functions specified in either (2) or (3), the SEPP is stationary if $K < 1$.

2.2 Estimation Methods for a SEPP

A standard estimation approach is to use maximum likelihood estimation (MLE). Let $\boldsymbol{\theta}$ be parameters in a SEPP, N, so $\boldsymbol{\theta} = (\eta, K, \beta)$ if the triggering function is exponential as in (2) and $\boldsymbol{\theta} = (\eta, K, \gamma)$ if the triggering function is power-law as in (3). Write $N(t) := N((0, t])$ and $0 = T_1 \leqslant T_2 \leqslant \cdots \leqslant T_n$ are n event times before time t where $N(t) = n$. The log-likelihood function for N based on observing $\{T_i\}_{i=1}^n$ (cf. Daley and Vere-Jones, 2003) is

$$l\left(\boldsymbol{\theta}; \{T_i\}_{i=1}^n\right) = -\int_0^t \lambda(s)\mathrm{d}s + \sum_{i=1}^{N(t)} \log(\lambda(T_i)), \tag{4}$$

and the conditional intensity function λ is calculated as in (1) with triggering function g that is either exponential (2) or power-law (3). Then the MLE is calculated as

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} := \mathrm{argmax}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}; \{T_i\}_{i=1}^n\right),$$

where the optimization can be done by the R function `optim` using methods such as `L-BFGS-B`. However, evaluating (4) takes $O(n^2)$ time and the log-likelihood can be almost flat in a large region of the parameter space; see Veen and Schoenberg (2008) for details. When applied to real data, the numerical optimization results are also sensitive to the choice of initial values of the model parameters. In Kirchner (2017); Kirchner and Bercher (2018), a faster estimation method is introduced. The main idea is to discretize time and approximate the count process on the resulting intervals by an integer-valued autoregressive model with finite order p, i.e. INAR(p). We now outline this estimation procedure.

Consider a SEPP N on a time window $[0, t]$. Choose a small Δ such that $0 < \Delta << t$ and some $p \in \mathbb{N}$ such that $p\Delta$ is large. This decomposes $[0, t]$ into roughly t/Δ bins. Here we do not choose too small a Δ as it will create too many zeros in the bin counts and therefore makes the estimation based on INAR(p) less reliable. See Kirchner (2017) for a discussion on the choice of p and Δ.

1. Discretization: calculate the number of edges added within each bin of length Δ to get the bin-count sequence $X_n^{(\Delta)}$, i.e.

$$X_n^{(\Delta)} := N\left(((n-1)\Delta, n\Delta]\right), \quad n = 1, 2, \ldots, [t/\Delta].$$

2. Optimization: find conditional LS estimates, i.e.

$$\left(\hat{\alpha}_0^{(\Delta)}, \ldots, \hat{\alpha}_p^{(\Delta)}\right) := \mathrm{argmin}_{(\alpha_0, \ldots, \alpha_p) \in \mathbb{R}^{p+1}} \sum_{n=p+1}^{[t/\Delta]} \left(X_n^{(\Delta)} - \alpha_0 - \sum_{k=1}^p \alpha_k X_{n-k}^{(\Delta)}\right)^2. \tag{5}$$

3. Normalization: Set,

$$\hat{\eta} := \frac{\hat{\alpha}_0^{(\Delta)}}{\Delta}, \quad \hat{K} := \sum_{k=1}^p \hat{\alpha}_k^{(\Delta)}, \quad \hat{g}_k^{(\Delta)} := \frac{\hat{\alpha}_k^{(\Delta)}}{\hat{K}\Delta}.$$

Note that $\hat{g}_k^{(\Delta)}$ is an estimator for $g(k\Delta)$. We assume the triggering function g belongs to one of the two parametric density families $(g_{\theta'})_{\theta' \in \Theta'}$ given in (2) and (3) and the unknown parameters are estimated from

$$\hat{\boldsymbol{\theta}} := \mathrm{argmin}_{\boldsymbol{\theta}' \in \Theta'} \sum_{k=1}^p \left(g\left((k-0.5)\Delta; \boldsymbol{\theta}'\right) - \hat{g}_k^{(\Delta)}\right)^2. \tag{6}$$

We use the R function `nls` using an algorithm from the `port` library to solve the optimization problem in (6).

The procedure above actually assumes the approximation,

$$\mathbf{E}(X_n^{(\Delta)}|\mathcal{H}_{(n-1)\Delta}) \approx \eta\Delta + K\Delta \int_{(n-p-1)\Delta}^{(n-1)\Delta} g(n\Delta - u)\mathrm{d}N(u)$$

and thus only observations within $((n-p-1)\Delta, (n-1)\Delta]$ have impacts on the triggering function $g(\cdot)$. The SEPP, therefore, has finite memory; see Jo et al. (2015); Zang et al. (2017) for more detail.

When applying the estimation procedure to real datasets in Section 3, we divide all timestamps into non-overlapping time windows (not to be confused with the bins of length Δ discussed above) with equal number of events and fit the SEPP model to each time window, rather than the whole dataset. While this exposes us to the dangers of overfitting, the reasons for this action are three-fold:

1. As the network evolves, there might be some change points in the evolution so that modeling the whole network by one simple model is not appropriate.
2. As noted in Kirchner (2017), if too many data points are included, then it is more likely to have abnormal observations which are not captured by the underlying model. Therefore, formal statistical tests will likely reject the model when assessing the goodness of fit for the SEPP model.
3. It may be that early in the life of the network, few events are observed which make fitting the model difficult. Hence, we set up the time windows by accumulating enough events, not by using a fixed period of time.

3 Network Data Examples

3.1 Facebook Wall Posts

The Facebook wall post dataset describes a regional network of users in New Orleans from 2004-09-13 to 2009-01-21. The data forms a directed graph where nodes are Facebook users and each directed edge represents a post from one node to another node's page. The dataset has three columns and 876,993 rows. The first two columns are anonymized user identifiers and the third column is a UNIX timestamp marking the time of the wall post. For each row, the second user posts on the first user's wall at the time given in the third column.

Figure 1 shows some summary information about the evolution of this regional network. The left graph gives cumulative number of edges (posts) in the network over time and the right panel displays the total number of new edges added per week. Both plots show that the network grows slowly during the first couple of years and after year 2006 the growth pattern becomes relatively stable. There is a sudden increase starting from mid-2008; Viswanath et al. (2009) speculate this is a consequence of Facebook's new site design unveiled on July 20, 2008 which allows users to directly view wall posts through friends' feeds.

Since the Figure 1 plots show a regular growth pattern after 2006, we start our analysis with the Facebook data before 2006-01-01. The point process prior to 2006-01-01 is more bursty than after, i.e. more clustered in time, and later we justify this by calculating the burstiness parameter B as defined below in (7).
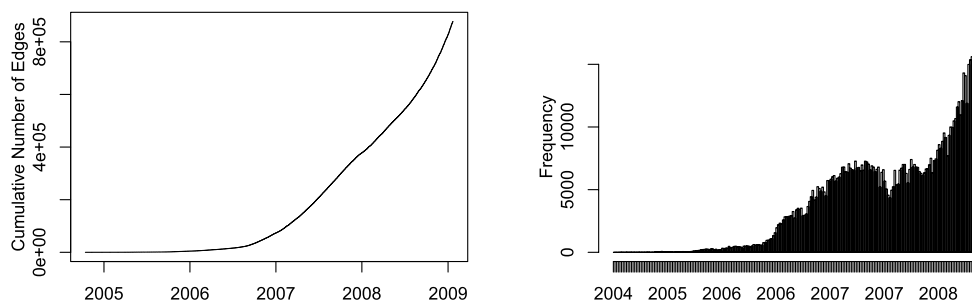
Figure 1: Facebook wall posts for users in New Orleans. Left: Cumulative number of edges over time. Right: number of new edges added per week.

Table 1: Burstiness parameters for IETs from 2005 to 2008, where B_{all} is calculated using all IETs and B_{active} is obtained by excluding all IETs that happen during 1–8 AM on each day. Since 2006, we see a decrease in the burstiness parameter after removing IETs during users' inactive hours.

	Yr 2005	Yr 2006	Yr 2007	Yr 2008
B_{all}	0.405	0.454	0.386	0.370
B_{active}	0.421	0.407	0.112	0.111

One way to measure the burstiness of a point process (Goh and Barabási, 2008) is through the *burstiness parameter*, B, defined as

$$B := \frac{\sigma - \bar{\tau}}{\sigma + \bar{\tau}}, \tag{7}$$

where σ and $\bar{\tau}$ are the standard deviation and mean of the inter-event times (IETs), respectively. In general, $B \in [-1, 1]$ and a highly bursty point process will have B close to 1 since σ is large relative to $\bar{\tau}$. For contrast, note that a homogeneous Poisson process has $B = 0$.

Calculating the burstiness parameter using all points prior to 2006-01-01 gives $B = 0.672$. The burstiness parameters for data after 2006 are summarized in the first row of Table 1 and the estimates are less than 0.672. The enhanced burstiness prior to 2006 suggests modeling the growth of the Facebook wall posts network by a SEPP, but we need to determine the proper triggering function. Since we only focus on the two choices in (2) and (3), we make the decision by assessing the goodness of fit under the two triggering functions.

To fit the SEPP, we convert the time scale from seconds to days and apply the estimation procedure in Section 2.2 (Kirchner and Bercher, 2018). We set $p = 150$, $\Delta = (1/24)/12$, which is a 5-minute time window, and fit the model locally to 55 disjoint time intervals each of which contains 200 events. Later we will apply a Kolmogorov-Smirnov (KS) test to assess the goodness of fit, but the KS test becomes very sensitive to even small deviations from the hypothesized distribution, when a large number of data points are included in one interval. As the network grows, we observe an increasing number of events within a window of fixed length, e.g. within one week. So here we will stick with a pre-specified number of events when fitting the model. Note that no observations from the previous interval are used while fitting the model. This covers the time period from 2005-01-01 to 2006-05-01. Due to the small number of observations

at the beginning of the Facebook network, we do not further decrease the value of Δ since it will generate too many zeroes, making the optimization setup in (5) less accurate. Also, choosing too large a p leads to error codes when doing computations in R so we do not further increase p.

Assuming a power law triggering function (3), we apply the method outlined in Section 2.2 and yielding the estimates

$$\tilde{\boldsymbol{\theta}} := \left(\tilde{\eta}, \tilde{K}, \tilde{\gamma} \right).$$

To assess the goodness of fit, we transform the event times

$$\Lambda(T_i; \boldsymbol{\theta}) := \int_0^{T_i} \lambda(t | \mathcal{H}_t; \boldsymbol{\theta}) \mathrm{d}t, \qquad i = 1, \ldots, n, \tag{8}$$

and evaluate $\{\Lambda(T_i) : i = 1, \ldots, n\}$ using parameters estimated by $\tilde{\boldsymbol{\theta}}$. With a power-law triggering function, estimated values of the transformed event times are

$$\Lambda(T_i; \tilde{\boldsymbol{\theta}}) := \tilde{\eta} \, T_i + \tilde{K} \sum_{k=1}^{i-1} \left(1 - (1 + T_i - T_k)^{-(\tilde{\gamma}-1)} \right). \tag{9}$$

If the model fit is good, the transformed inter-event times,

$$\{\Lambda(T_i; \tilde{\boldsymbol{\theta}}) - \Lambda(T_{i-1}; \tilde{\boldsymbol{\theta}}) : i = 1, \ldots, n\} \tag{10}$$

should be iid exponentially distributed with unit rate (cf. Proposition 7.4.IV in Daley and Vere-Jones, 2003). Examining the plots of autocorrelation function (ACF) (collected in the online supplement on https://github.com/tw398/NetworkGrowth) shows little autocorrelation among transformed IETs. Then we applied a KS test to see whether the transformed IETs calculated as in (9) and (10)) follow an exponential distribution with unit rate. Overall, we think the model fit is adequate if the KS test returns an approximately 50% acceptance rate with level $\alpha = 0.01$ over all time windows. It turns out that the KS test does not reject the SEPP model with a power-law triggering function (3) for only 1 out of the 55 disjoint time windows, no matter whether the significance level, α, is set to 0.01 or 0.05. Moreover, estimates of the index of the power-law function, γ, are large (with a median ≈ 28.6 based on the 55 estimates), which indicates that the decay rate of the triggering function is fast. In order to look for common growth patterns across different datasets, we then restrict ourselves to an exponential triggering function as in (2), which is widely applied in the literature; see for example Rizoiu et al. (2017a).

Therefore, we refit the SEPP model to the same 55 disjoint time intervals but this time used the exponential triggering function. Setting $p = 150$, $\Delta = (1/24)/12$ and applying the method outlined in Section 2.2 give a new set of estimates:

$$\widehat{\boldsymbol{\theta}} := \left(\widehat{\eta}, \widehat{K}, \widehat{\beta} \right). \tag{11}$$

Again, we calculate the transformed times in (8) using $\widehat{\boldsymbol{\theta}}$ and test the empirical distribution of the transformed IETs against the hypothesis of an exponential distribution with unit rate. With an exponential triggering function (2), the transformed event times defined in (8) become

$$\Lambda(T_i; \widehat{\boldsymbol{\theta}}) := \widehat{\eta} \, T_i + \widehat{K} \sum_{k=1}^{i-1} \left(1 - e^{-\widehat{\beta}(T_i - T_k)} \right), \qquad i = 1, \ldots, n, \tag{12}$$

and the transformed IETs are

$$\{\Lambda(T_i; \widehat{\boldsymbol{\theta}}) - \Lambda(T_{i-1}; \widehat{\boldsymbol{\theta}}) : i = 1, \dots, n\}. \tag{13}$$

From the ACF plots (see supplement), we do not observe large autocorrelation among transformed IETs and the acceptance rate of the KS test increases to 48.1% with $\alpha = 0.05$. The rate reaches 64.8% when $\alpha = 0.01$. Therefore, we conclude that the SEPP with an exponential triggering function provides a better fit to the Facebook data from 2005-01-01 to 2006-05-01. Note that due to the small number of events at the beginning of the network, it takes a long time to accumulate 200 consecutive events. In fact, the actual durations of the first 10 time windows are all longer than a week. For a long time duration, it is possible to observe abnormal IETs which leads to rejections by the KS test. If we remove the first 10 windows, the two acceptance rates increase to 50.0% and 67.4%, respectively. Accounting for such fact, we think the obtained acceptance rate indicates a good fit of the SEPP model.

The analysis is then extended to the next 100 disjoint time windows each of which includes 200 points and this covers the time period from 2006-05-01 to 2006-09-22. However, the estimation algorithm in Section 2.2 fails when applied to data subsequent to the 57-th time window, indicating a change in the underlying model should be made. Looking at the first 56 time windows (from 2006-05-01 to 2006-08-20), we see that the acceptance rates of the KS test decrease to 40% with $\alpha = 0.05$ and 56.4% with $\alpha = 0.01$. Since we see decreases in the acceptance rates and a failure of the estimation algorithm from 57-th time window onwards, we conclude that 2006-05-01 to 2006-09-22 is a period where a model change seems to take place.

Figure 2 summarizes the parameter estimates $\widehat{\boldsymbol{\theta}}$ (cf. (11)) over the total 110 disjoint time windows covering the period from 2005-01-01 to 2006-08-20. Within each time window, 200 points are included. We notice that after the 64-th time window, $\widehat{K}$ starts to take negative values, which deviates from the model assumption in (2). Moreover, after the 64-th time window, negative values of $\widehat{K}$ are offset by large positive values of $\widehat{\eta}$. For the two negative incidences of $\widehat{K}$ before the 64-th time window, the magnitude of $\widehat{K}$ is small (very close to 0) and the KS test does not accept the model. Therefore, we conclude that changing the underlying model is necessary from the 64-th time window onward (i.e. after 2006-05-24) and the SEPP with an exponential triggering function can only be used to model the Facebook data from 2005-01-01 to 2006-05-23.
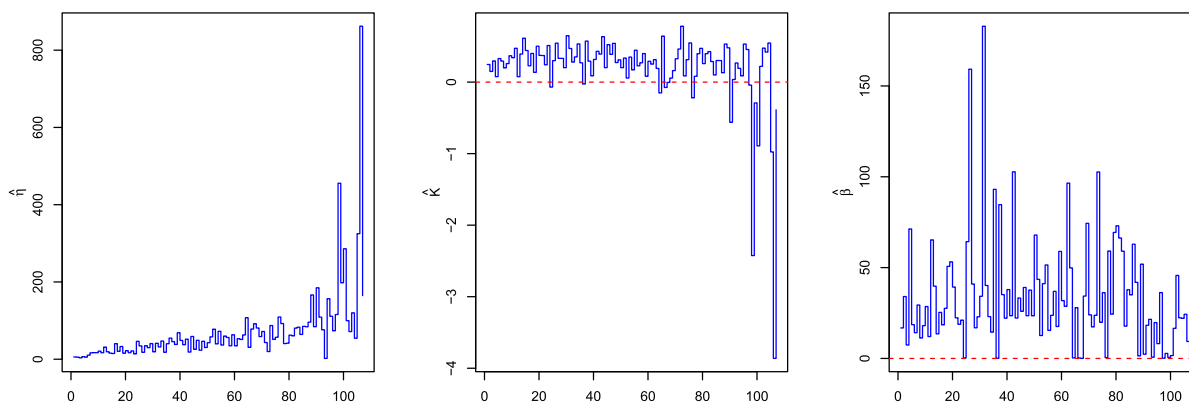

Figure 2: Estimated values of $\widehat{\boldsymbol{\theta}} = \left(\widehat{\eta}, \widehat{K}, \widehat{\beta}\right)$ for the Facebook data from 2005-01-01 to 2006-08-20. Each set of parameter estimates is calculated using 200 points. The model struggles to fit after the 64-th time window.
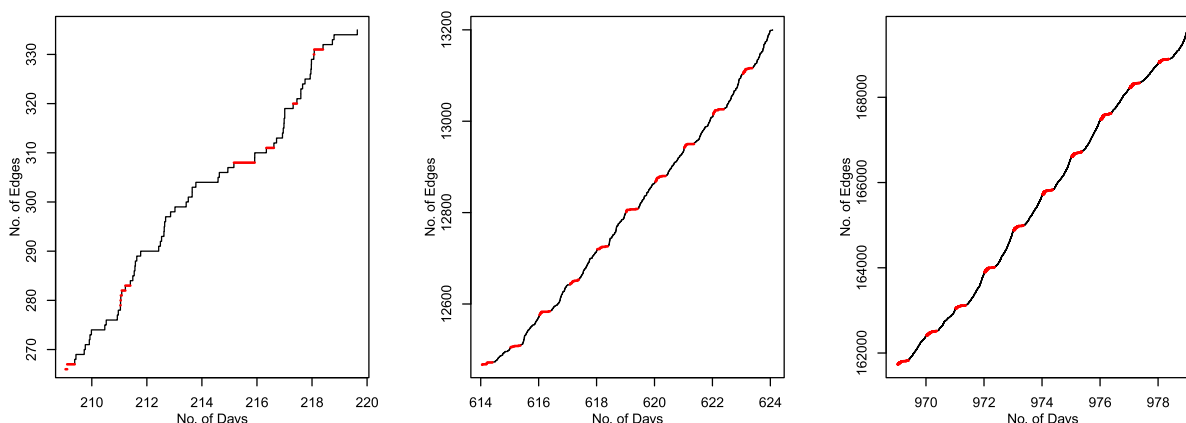
Figure 3: Cumulative number of edges in two periods: (1) Left: 2005-04-11 to 2005-04-20 (when the SEPP model fits well); (2) Middle: 2006-05-21 to 2006-05-30 (when the SEPP model fits poorly). (3) Right: 2007-05-11 to 2007-05-20. Red line segments refer to edges created during 1–8 AM each day (US Central Time Zone), when we expect sleep interferes with Facebook activity. The left plot looks more bursty and erratic than the other two and a repetitive growth pattern is observed in the middle and right plots.

When inspecting the network growth pattern after 2006-05-24, we see a repeating pattern. For comparison, in Figure 3 we plot the cumulative number of edges from 3 representative periods:

1. 2005-04-11 to 2005-04-20, where the SEPP model fits well.
2. 2006-05-21 to 2006-05-30, where the SEPP model starts to give poor fit to the data indicating a change in the model.
3. 2007-05-11 to 2007-05-20, where a common repetitive pattern is observed.

The line segments in red correspond to the accumulation of new edges from 1 to 8 AM each day, during which we expect sleep interferes with Facebook activity, and the timestamps are adjusted according to the US Central Time Zone. When no accumulation of edges has been made during 1–8 AM, then we do not color the sleeping hours as red.

In a plot of cumulative number of edges, a relatively long flat line segment corresponds to a relatively long IET, meaning a long period of inactivity in the network. Comparing the three panels in Figure 3 shows that at first when the SEPP model is accepted, the longer IETs are not necessarily associated with anticipated sleeping periods but later on longer IETs are usually associated with the expected inactive hours (during 1–8 AM). Moreover, Table 1 summarizes the burstiness parameter based on the point process from 2005 to 2008, where B_{all} is calculated using all timestamps during the year and B_{active} excludes timestamps that occur during 1–8 AM each day. Note that from 2006 onward, excluding timestamps during expected sleeping periods makes the point process less bursty.

Next we exclude timetamps that occur during 1–8 AM on each day and obtain reduced data sets for each day. Motivated by the observed decrease in the burstiness of the process shown in Table 1, we fit a NHPP to the reduced point process with constant rate across a day but changing rates from day to day. The estimated rate parameter of the Poisson process is calculated using MLE and this procedure is applied to data from 2006-05-01 to 2008-12-31. Estimation results are given in Figure 4.
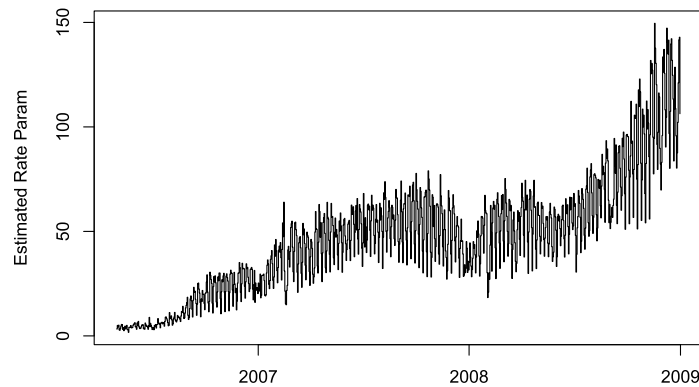
Figure 4: MLE's of the rate parameter when a homogeneous Poisson process is fitted to daily cleaned IETs (excluding IETs during 1–8 AM) for the Facebook data from 2006-05-01 to 2008-12-31.

We then test each day's reduced IET data against the exponential null. We also assess our exponential fit by conducting a KS test. For $\alpha = 0.01$, 90.3% of the time the KS test does not reject the exponential fit. With $\alpha = 0.05$, the acceptance rate of the KS test is 76.2%. Meanwhile, from mid 2007 to mid 2008, the estimated Poisson rate remains relatively stable but keeps increasing after mid 2008, which matches what we have observed from Figure 1.

3.1.1   Summary of Facebook Wall Posts Data

From the analysis of the Facebook wall posts data, we see that the network growth of the Facebook wall posts (New Orleans) has three different phases.

1. Prior to the first quarter of year 2006, the point process is bursty and we model the network growth by a SEPP with an exponential triggering function as specified in (2). We consider this time period as the startup phase of the network.
2. Starting from May 2006, we observe a repetitive growth pattern. Users are less active on the network during 1–8 AM when we expect users are asleep. After removing timestamps occurring between 1–8 AM, the reduced point process becomes less bursty and can be modeled by a NHPP where the Poisson rate is constant across a day but changing as we move from day to day.
3. From mid 2007 to mid 2008, the daily estimates of the Poisson rate parameter come to a relatively stable phase but start increasing again after mid 2008, perhaps as suggested by Viswanath et al. (2009) as a result of the release of Facebook's new design.

3.2   Dutch Wiki Talk

We now discuss data from another regional social network, the Dutch Wikipedia Talk. This dataset has been analyzed in Wan et al. (2017) where we fit a linear preferential attachment model, but here we only consider the timestamp information in this dataset. We take the cleaned data from Wan et al. (2017) where inferred administrative broadcasts and posts are removed and look at remaining edges created from 2003-01-01 to 2013-12-31. The data forms a directed network where nodes are users and edges denote user interactions on the talk pages in Dutch Wikipedia.
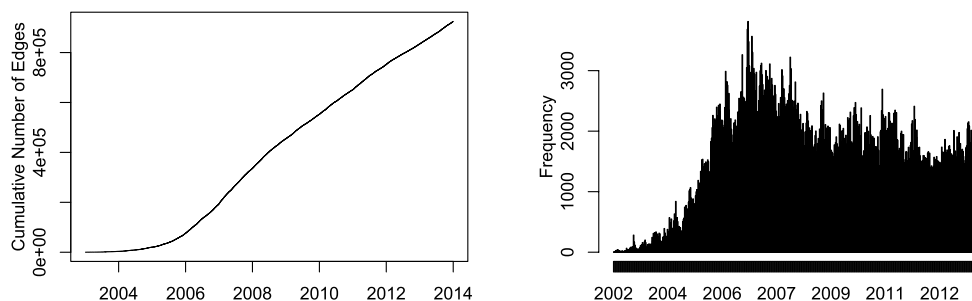
Figure 5: The evolution of the Dutch Wiki Talk data. Left: Cumulative number of edges over time. Right: Total number of new edges added per week.

This dataset contains 924,473 rows and three columns. The first two columns are users' anonymized ID's and the third column is a UNIX timestamp for the time when the user in column 1 writes a message on the Wiki talk page of the user in column 2. The summary of timestamps for this network is in Figure 5. We adjust all timestamps according to the Central Europe Time Zone. Different from the Facebook wall post data, the Dutch Wiki Talk data displays an accelerating growth between 2003 and 2007 but starting from 2008 the growth pattern is close to linear.

Similar to the Facebook case, after accounting for the poor fit of a SEPP model with a power-law triggering function (3), we settle on the exponential triggering function (2). We divide the Dutch Wiki talk data from 2003-01-01 to 2005-12-31 into 151 disjoint time windows, each of which contains 500 events, and fit the SEPP model to each time window using the algorithm of Section 2.2 with $p = 150$ and $\Delta = (1/24)/12$. We denote estimated model parameters by $\widehat{\boldsymbol{\theta}}$ with three components as detailed in (11). With $\widehat{\boldsymbol{\theta}}$, we compute the transformed time $\Lambda(T_i)$, $i = 1, 2, \ldots, n$ as in (8) and use a KS test to examine whether the transformed IETs (calculated using (12) and (13)) are exponentially distributed with rate 1. Out of the 151 sets of estimates, the acceptance rate for the KS test is 30.7% with $\alpha = 0.05$ and 46.7% with $\alpha = 0.01$. Also, the ACF plots (see supplement) do not suggest high autocorrelation among the transformed IETs.

This acceptance rate looks lower than for the Facebook case but now 500 events are included in each estimation interval as opposed to 200 events for Facebook. With a larger sample size, we are more likely to have abnormal IETs that deviate from our model and hence reject the null. Meanwhile, the Dutch Wiki case also has a similar problem as for Facebook. At the beginning of the network recording period, it takes a long time to accumulate 500 events (10 days or more for the first 22 time windows), which again increases the possibility of having abnormal observations. The bigger sample size, slow network startup and the sensitivity of the KS test to outliers encourage us to proceed assuming the SEPP model for this time period.

Estimated values of parameters are given in Figure 6, where the left, middle and right panels give $(\widehat{\eta}, \widehat{K}, \widehat{\beta})$, respectively. From the left plot of Figure 6, we observe a positive trend in $\widehat{\eta}$ with larger variation after 100 time windows (2005-09-07). Meanwhile, $\widehat{\beta}$ fluctuates more after the 46-th time window (2005-02-09), and the large magnitudes of $\widehat{\beta}$ suggest a fast decay of the self-excitation. Also, $\widehat{K}$ jumps out of the stationary range $(0, 1)$ after 100 time windows. When $\widehat{K}$ is negative, it is compensated by a large positive $\widehat{\eta}$. These facts indicate the model increasingly struggles to fit the data as time progresses.

We further extend the estimation procedure to the data from 2006-01-01 to 2006-12-31, namely 236 disjoint time windows each of which has 500 points. The same KS tests are applied
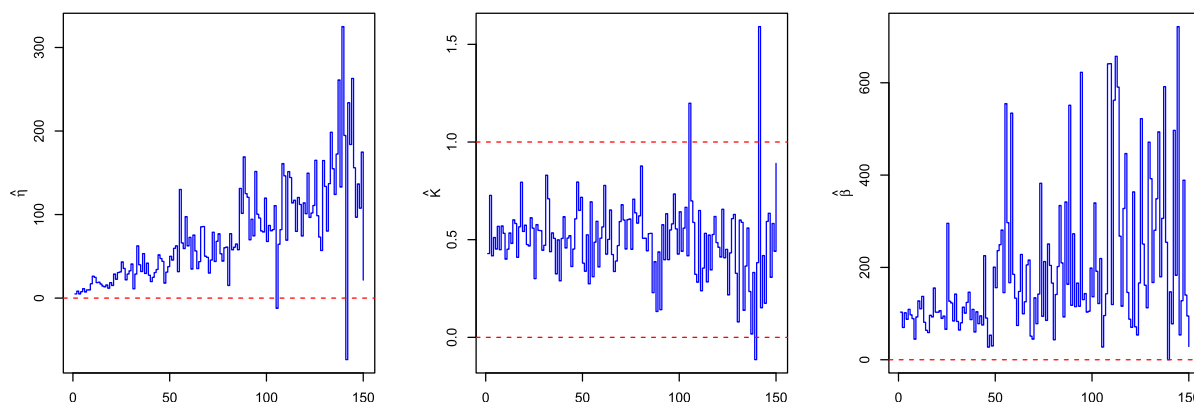
Figure 6: Left to right: Estimates for η, K, β from the Dutch Wiki talk data 2003-01-01 to 2005-12-31. Each set of estimates are computed using 500 points. After 45 disjoint time windows, estimates for β are more volatile and after 100 windows estimates for K occasionally jump outside of the $(0, 1)$ range.
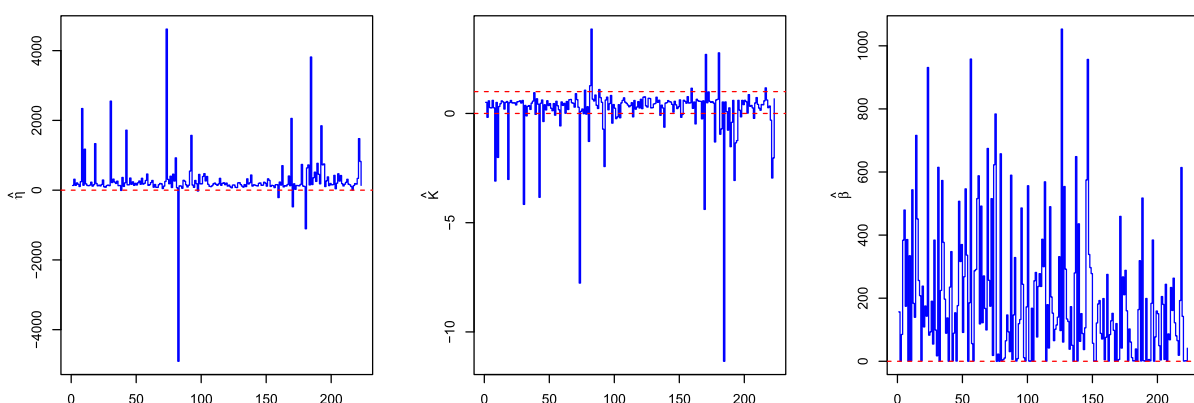

Figure 7: Left to right: Estimates for η, K, β based on the Dutch Wiki talk data from 2006-01-01 to 2006-12-31. Each set of estimates are computed using 500 points. The three estimated parameters are quite variable. A large positive $\widehat{\eta}$ is associated with a negative $\widehat{K}$, which indicates problematic fit.

to transformed IETs and now with $\alpha = 0.05$, the acceptance rate decreases to 19.7% and 31.4% when $\alpha = 0.01$. Estimated parameter values are plotted in Figure 7. We see from Figure 7 that $\widehat{\eta}$ fluctuates more and a large positive $\widehat{\eta}$ is likely to be accompanied by a negative $\widehat{K}$. Estimated values for β are also volatile and sometimes become very close to zero. Therefore, we are uncomfortable with the fit and seek a different model for data after 2005-02-09.

Looking at data after 2005-02-09, similar to the Facebook case, we also find a similar repetitive pattern. Figure 8 gives an example where we plot the cumulative number of edges created from

1. 2003-10-11 to 2003-10-18: the SEPP model fits well.
2. 2005-02-07 to 2005-02-13: the SEPP model starts to fit poorly.
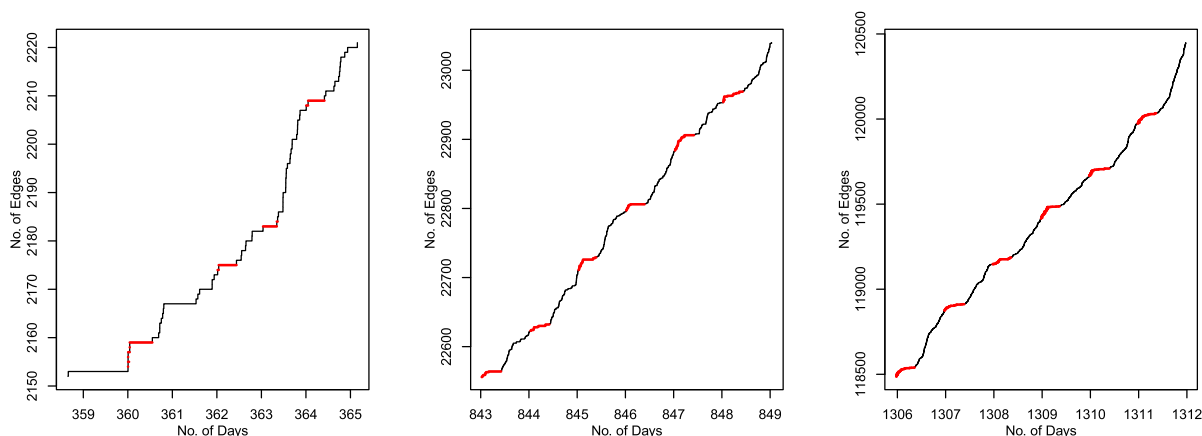3. 2006-05-16 to 2006-05-22: a clear repetitive pattern is observed.

Figure 8: Cumulative number of edges during three periods. Left: 2003-10-11 to 2003-10-18. Middle: 2005-02-07 to 2005-02-13. Right: 2006-05-16 to 2006-05-22. Red line segments refer to edges created during 0–9 AM each day (Central Europe Time Zone), when we expect sleep interferes with Dutch Wiki talk activity. We observe a repeating pattern in the network growth in the middle and right panels.

Table 2: Burstiness parameters for IETs of the Dutch Wiki talk data from 2003 to 2007, where B_{all} is computed using all IETs and we calculate B_{active} by excluding all IETs occurring during 0–9 AM on each day. Excluding IETs from 0–9 AM decreases the burstiness parameter.

	Yr 2003	Yr 2004	Yr 2005	Yr 2006	Yr 2007
B_{all}	0.441	0.395	0.460	0.447	0.471
B_{active}	0.448	0.401	0.375	0.257	0.213

All IETs associated with edges created between 0–9 AM each day are marked as red segments in Figure 8. We see that after the SEPP model starts to return a poor fit in 2006, longer IETs are often associated with the time period when people usually sleep and are therefore inactive on networks.

Table 2 summarizes the burstiness parameters where B_{all} is computed using all timestamps and we calculate B_{active} by excluding timestamps occurring during 0–9 AM on each day. We see that the point process is progressively less bursty after timestamps between 0–9 AM are removed.

Based on the evidence of decreasing burstiness in Table 2 and also Figure 8, we exclude timestamps created between 0–9 AM on each day after 2005-02-01 and fit a NHPP to the cleaned data with a constant rate within each day. We plot estimated daily rate parameters in Figure 9 based on data from 2005-02-01 to 2013-12-31. The Poisson rate estimates are increasing until 2008 and remain relatively stable from 2008 onward.

We use a KS test for a null hypothesis of exponentiality applied to the daily cleaned data. We summarize the acceptance rate of the KS test in Table 3 with $\alpha = 0.01, 0.05$. We see that the acceptance rate of the KS test is increasing over each year indicating an improvement in the Poisson fit. Hence, for the Dutch Wiki talk network from 2005 to 2008, we can model the daily point process without the points during 0–9 AM as a NHPP but the estimated rate parameters
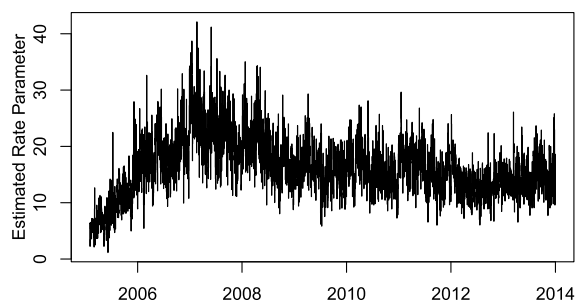
Figure 9: MLE's of the exponential rate parameter when a NHPP is fitted to daily cleaned IETs (excluding IETs during 0–9 AM) based on the Dutch Wiki talk data from 2005-02-01 to 2013-12-31.

Table 3: Acceptance rate (%) of the KS test for a NHPP model fitted to the Dutch Wiki talk data from 2005-02-01 to 2013-12-31. We assume the rate parameter is constant across the day but varies from day to day. The fit improves since year 2008.

Yr 20XX	05	06	07	08	09	10	11	12	13
$\alpha = 0.05$	30.9	22.7	27.9	31.2	34.5	36.7	35.1	35.1	32.3
$\alpha = 0.01$	51.5	40.0	49.6	49.3	60.0	60.3	58.6	62.7	55.1

are in general increasing. After 2008, the Poisson fit to the cleaned data is good and the estimated rate parameters stay relatively stable.

3.2.1 Summary of Dutch Wiki Talk Data

From the analysis of the Dutch Wiki talk data, we conclude that the growth pattern has three phases:
1. In the startup phase of the network until early 2005, the point process is bursty and well-modelled by a SEPP with an exponential triggering function (2).
2. From February 2005 onward, after accounting for a daily inactive cycle from we model remaining timestamps by a NHPP with constant rate across a day but changing rates from day to day, Noticeably from Figure 9, the daily rates fluctuate more and keep increasing from early 2005 to 2007, which can be regarded as the growing phase. Large fluctuations may be due to abnormal observations in the data, and the large fluctuations may account for the lower acceptance rate from the KS test; see Table 3.
3. Starting from 2008, the Poisson rate estimates (after removing timestamps between 1–9 AM) become relatively stable, and we think of it as the stable phase of the network growth.

In the interests of parsimony, one could try to fit an intensity function to the plot in Figure 9 of the NHPP model used in the growing and stable phases. For instance, by looking for common trends in the rate estimates within a week. We give a brief discussion later in Section 4.2. However, we have not pursued this carefully.
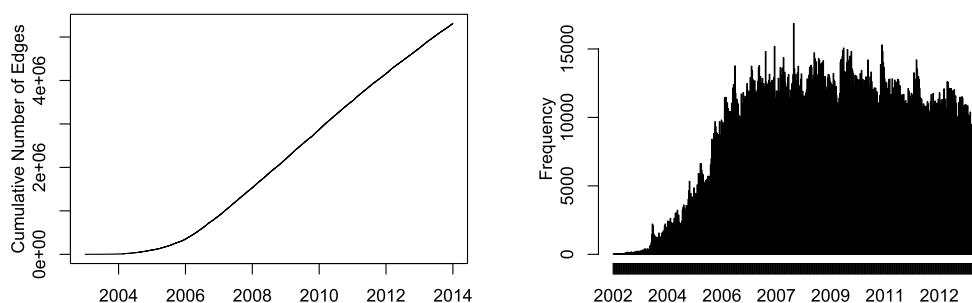
Figure 10: The evolution of the German Wikipedia Talk network from 2003-01-01 to 2013-12-31. Left: Cumulative number of edges. Right: The number of edges added per week.
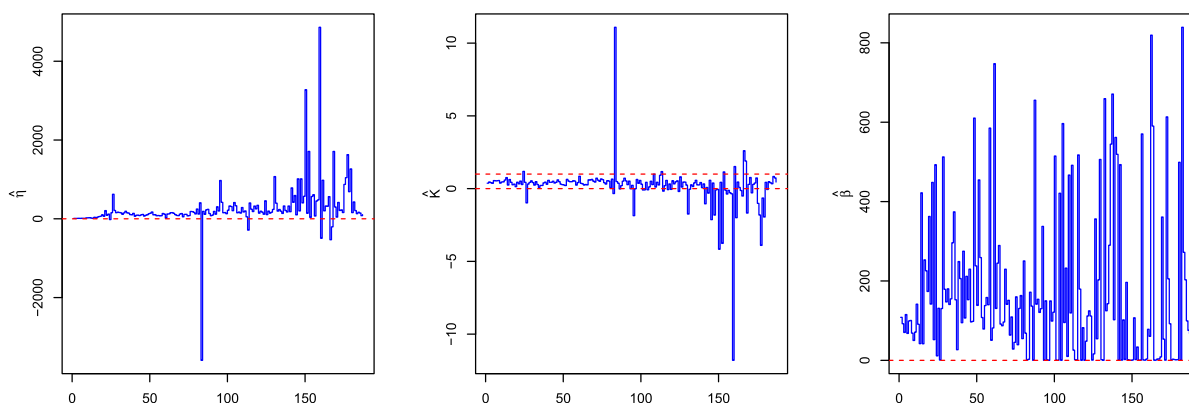

Figure 11: Left to right: Estimated values for $\widehat{\eta}$, $\widehat{K}$, $\widehat{\beta}$. Note that $\widehat{\eta}$ fluctuates more after the 82-nd time window (2004-07-19) and a large $\widehat{\eta}$ is offset by a negative $\widehat{K}$, indicating a problematic fit.

3.3 German Wiki Talk

In this section, we consider the Wikipedia Talk data in German, which takes a similar format as in the Dutch case, but the dataset is much larger. We first clean the data by removing all edges created by "administrators" and "bots". This is done by Wiki-talk datasets (2016), where users of the German Wikipedia are classified into three different groups, namely "normal users", "administrators" and "bots". We again look at the data from 2003-01-01 to 2013-12-31. The cleaned dataset contains 5,863,373 rows and three columns. No doubt the larger size of German Wikipedia compared to Dutch Wikipedia is a consequence of the much larger number of fluent German speakers (90–95 million according to German Wikipedia, 2019) compared to Dutch speakers (29 million by Dutch Wikipedia, 2019).

Figure 10 displays summary timestamp plots where the cumulative number of edges over time is plotted in the left panel and the number of new edges added per week is given in the right panel. All timestamps are adjusted to the Central Europe Time Zone. The shape of the graph looks very similar to the Dutch case, but the German Wiki Talk data grows faster.

We start by examining the data from the first two years, namely from 2003-01-01 to 2004-12-31, and divide the data into 196 disjoint time windows each of which includes 500 points. Fitting a SEPP model with an exponential triggering function (using the method in Section 2.2 with $p = 150$ and $\Delta = (1/24)/12$) to data in these 196 time windows gives estimates of $\widehat{\theta}$
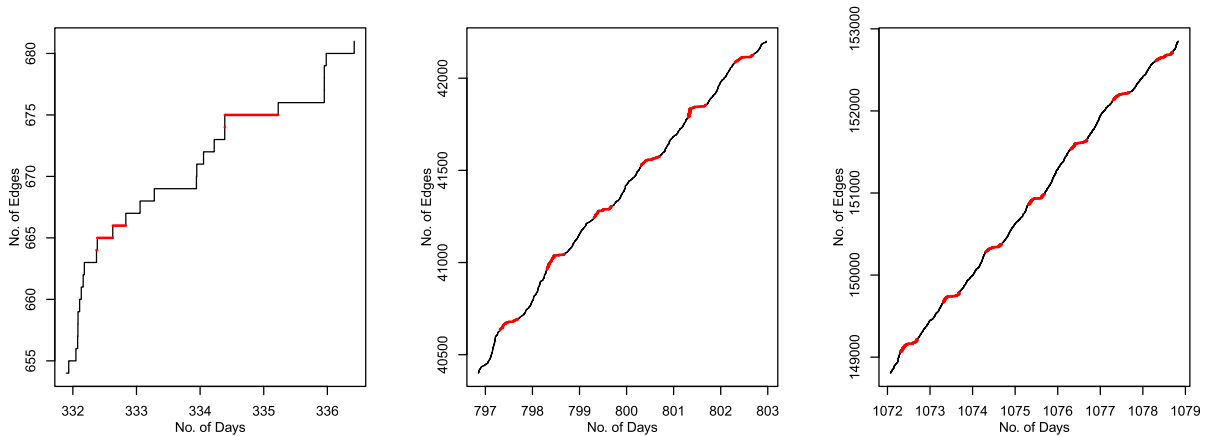
Figure 12: Cumulative number of edges during three periods. Left: 2003-04-06 to 2003-04-12. Middle: 2004-07-15 to 2004-07-21. Right: 2005-04-15 to 2005-04-21. Red line segments give edges created during 0–9 AM each day (Central Europe Time Zone), when sleep interferes with German Wiki talk activity. Network growth has a repeating pattern in the middle and right panels.

as in Figure 11. Subsequent to the 82-nd time window (2004-07-19), $\widehat{K}$ becomes more volatile and often a negative $\widehat{K}$ is offset by a large positive value of $\widehat{\eta}$. Fluctuations of the estimates $\widehat{\beta}$ also increase after the 82-nd time window and sometimes the estimated value is close to zero. These observations from Figure 11 suggest the SEPP model struggles to fit the data after the 82-nd time window. In fact, the acceptance rate of the KS test which examines whether the transformed IETs (calculated using (12) and (13)) are exponentially distributed with unit mean is 25.1% when $\alpha = 0.05$ and 34.2% when $\alpha = 0.01$, if all 196 time windows are taken into consideration. However, if only focusing on the first 81 time windows, the acceptance rate of the same KS test increases to 36.6% for $\alpha = 0.05$ and 47.6% for $\alpha = 0.01$. At the beginning of the time record, longer time periods are required to accumulate 500 events (10 days or more for the first 16 time windows) and so we are unlikely to obtain good fits and in light of this observation we conclude that the SEPP model gives an acceptable fit for the network growth until July 2004. We change the model for data after 2004-07-19.

Similar to the Dutch Wiki case, a repeating pattern is observed in the growth of the German Wiki talk network. Figure 12 plots the cumulative number of edges during three time periods:
1. 2003-04-06 to 2003-04-12: the SEPP model fits well.
2. 2004-07-15 to 2004-07-21: the SEPP model fits poorly, suggesting changes in modeling.
3. 2005-04-15 to 2005-04-21: the point process exhibits a repetitive pattern.

Red line segments correspond to edges created during 0–9 AM on each day. From the middle and right panels of Figure 12, as expected, we see a repeating pattern and users are less active from 0 to 9 AM on each day. This agrees with the findings in the Dutch Wiki case.

For the data from 2004-07-01 to 2013-12-31, we remove all timestamps during 0–9 AM and fit a NHPP model to the cleaned data such that the rate parameters are constant across the day but vary from day to day. The acceptance rate for the KS test assessing the Poisson fit is reported in Table 4. In general, fits after 2005 are better than before and on average, the acceptance rate is 35.9% when $\alpha = 0.05$ and 58.2% when $\alpha = 0.01$. The estimated rate parameters (using reduced daily data) are plotted in Figure 13. The rate estimates exhibit increasing trend until roughly 2007 and stay relatively stable afterwards. Estimates fluctuate more during 2007–2010, which

Table 4: Acceptance rates (%) of the KS test for a NHPP model fitted to the German Wiki talk data from 2004-07-01 to 2013-12-31. The fit improves subsequent to year 2005. On average, the acceptance rate is 35.9% when $\alpha = 0.05$ and 58.2% when $\alpha = 0.01$.

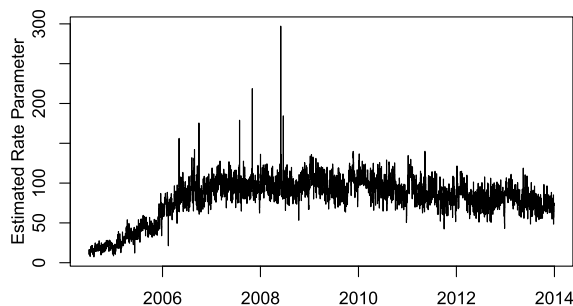| Yr 20XX	04	05	06	07	08	09	10	11	12	13
$\alpha = 0.05$	21.1	51.5	41.1	20.2	26.3	27.1	27.4	44.7	40.8	41.1
$\alpha = 0.01$	33.7	73.4	63.0	41.1	50.4	49.3	48.8	67.9	66.3	59.2

Figure 13: MLE's of the rate parameter when a NHPP is fitted to daily cleaned IETs (excluding IETs during 0–9 AM) based on the German Wiki talk data from 2004-07-01 to 2013-12-31.

is associated with lower acceptance rates of KS test in Table 4. This may be due to abnormal IETs during that particular time period which cannot be caught by the Poisson model.

3.3.1 Summary of German Wiki Talk Data

Conclusions for the German Wiki data are similar to the Dutch case. The network is in startup phase until July 2004 during which the growth can be modeled by a SEPP with an exponential triggering function. After the startup phase ends, edge creation can be viewed as a NHPP with constant rate within a day but different rates from day to day, plus a daily inactivity time period during which users are presumably asleep.

The difference between the German Wiki data and the Dutch one is the growth rate of the network. Although both were founded in 2001 (Dutch Wikipedia, 2019; German Wikipedia, 2019), German Wikipedia has more articles and the total number of German speakers (90–95 million) is also much larger than that of Dutch speakers (29 million) (Dutch Wikipedia, 2019; German Wikipedia, 2019). These may accelerate the growth of the network and shrink the length of the startup phase.

3.4 Math Overflow

After seeing three regional communication networks that are geographically localized, for contrast we investigate another network which is not regional. The MathOverflow data includes interactions from the "StackExchange" site "MathOverflow". The network was launched September 28, 2009. Types of interactions on MathOverflow include answering a question from another user, commenting on another user's question and commenting on another user's answer. In the graphical representation of this network, users are nodes and online interaction between two users is regarded as a directed edge. The dataset contains 506,550 edges (rows) and covers the
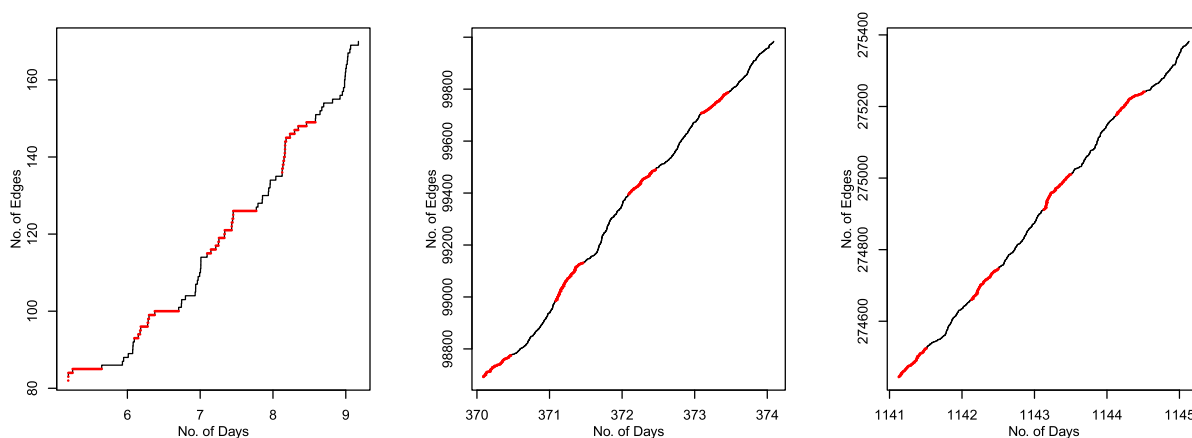
Figure 14: Cumulative number of edges from the MathOverflow data during three representative time periods: (1) 2009-10-03–2009-10-07 (left; the beginning of the network); (2) 2010-03-02–2010-03-06 (middle; about one year after the launch of MathOverflow); (3) 2012-11-12–2012-11-16 (right; about two years after the launch of MathOverflow). Red line segments refer to edges created during 0–9 AM each day, based on US Central Time. Unlike the regional networks, the repeating pattern with a nightly inactivity period is not observed here, perhaps because the MathOverflow website is used by users worldwide and thus is not regional.

time period from 2009-09-28 to 2016-03-06. There are three columns in the dataset. The first two columns represent users' ID and the third column is a UNIX timestamp recording the time of the interaction. For each row, the first user interacts with the second user in one of the three ways given above.

Since the MathOverflow website can be accessed by users worldwide, there is no obvious choice of a particular time zone to adjust the timestamps associated with the creation of each edge. We simply set the time zone to be the US Central Time and Figure 14 plots the cumulative number of edges within three time periods:

1. 2009-10-03–2009-10-07, which is the very beginning of the network.
2. 2010-10-03–2010-10-07, approximately one year after the network has been launched.
3. 2012-11-12–2012-11-16, approximately two years after the network has been launched.

Comparing Figure 14 with Figure 3, 8 and 12, we see that the daily repeating pattern vanishes as the network evolves and no nightly period of inactivity is evident.

We try fitting a SEPP model to the first 5 months of the network data (from 2009-10-01 to 2010-02-28) by creating 69 consecutive time windows each of which contains 500 points. Estimation results are presented in Figure 15. Then we again use a KS test to compare the transformed IETs (calculated using (12) and (13)) with the null hypothesis of an exponential distribution with unit mean. Although the acceptance rate of the KS test is 42.0% when $\alpha = 0.05$ and 58.0% when $\alpha = 0.01$, 15 out of the 69 $\widehat{K}$ values (the middle panel of Figure 15) are negative. When inspecting the ACF plots for transformed IETs (see supplement), large autocorrelations raise concern about the goodness of fit of the SEPP model. So even though the KS test has a reasonable acceptance rate, we conclude the SEPP is not suitable to model early-stage growth of MathOverflow.

Given that no clear repetitive daily inactivity period has been observed, we try fitting a NHPP to the timestamps such that IET data within one day is tested against an exponential
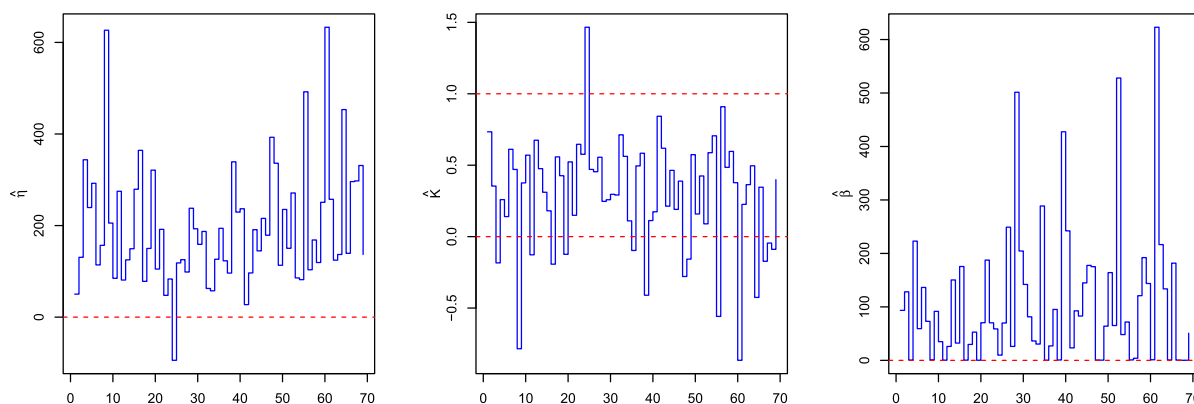
Figure 15: Estimated values for $(\widehat{\eta}, \widehat{\beta})$ (left) and $\widehat{K}$ (right) based on the MathOverflow data from 2009-10-01 to 2010-02-28. Each set of estimates is computed using 500 points. Note that 15 out of 69 estimated $\widehat{K}$ values are negative.

Table 5: Acceptance rate (%) of the KS test for a NHPP model fitted to the MathOverflow data from 2010-01-01 to 2015-12-31. The acceptance rate is high, but autocorrelation between IETs is not negligible.

Yr 20XX	10	11	12	13	14	15
$\alpha = 0.05$	67.1	67.7	76.4	70.7	66.3	65.5
$\alpha = 0.01$	84.9	88.2	92.3	86.3	85.2	87.9

null with a constant rate, but the rate parameter varies from day to day. The rate parameter of the Poisson process is estimated using MLE. The acceptance rates of the KS test for the MathOverflow data from 2010-01-01 to 2015-12-31 are summarized in Table 5. The acceptance rates stay very high for each year but at the same time the autocorrelation among IETs (ACF plots are given in the supplement) is also significant, indicating more analysis could be conducted to try to account for the correlation.

Our conclusion is that the SEPP + NHPP model is not a good fit for the MathOverflow data. The growth pattern for this network which is not geographically concentrated is different from the three regional networks we looked at.

4 Discussion

4.1 Common Growth Pattern

We reviewed four social network datasets. Three of the networks are geographically concentrated, namely Facebook wall posts for users in New Orleans, Dutch and German Wikipedia talk messages. The MathOverflow data does not have the geographical concentration property. The regional social networks possess common evolution patterns that are not shared by Math-Overflow:

1. Startup phase: When the network is building up, network evolution can be modeled by a SEPP with an exponential triggering function. During this startup period, longer IETs are

not necessarily associated with daily inactive periods for users. Meanwhile, the length of the startup phase varies and may depend on the number of users of the network. For instance, the startup phase disappears within roughly 1.5 years for the German Wiki data, after which the SEPP model fits poorly. In contrast, the startup phase in the Dutch Wiki case lasts for more than two years. The different startup lengths is presumably a consequence of the difference in the number of possible users of the network. The total number of German speakers worldwide is about 90–95 million whereas that of Dutch speakers in the world is around 29 million and more available users have the potential to accelerate the growth of the network, making the startup phase disappear more quickly.

2. Growing phase: After the network has attracted some attention, it enters a growing phase and the daily evolution can be modeled by a nightly inactive period plus daytime activity described by a NHPP, where the Poisson rate changes from day to day but stays constant across each day, The estimated rate parameters of the Poisson process while not monotone are trending upwards during the growing phase.

3. Stable phase: The regional network becomes relatively saturated and the overnight inactivity persists. After reaching the stable phase, the rate parameters from the Poisson process remain stable unless some new external stimulus is brought to the network. For example, in the Facebook case, we observe a sudden increase in the estimated rate parameters since mid-2008, which may be due to Facebook's release of a new site design on July 20, 2008 (Viswanath et al., 2009).

This three-stage growth pattern does not apply to networks that are not geographically concentrated, as exemplified by the MathOverflow data. Although the acceptance rate of a Poisson model is high, the serial dependence among IETs is not negligible and should be accounted for.

We also realize that there are some limitations in the current study. First, the SEPP models fitted here assume constant baseline intensity over each window, and this assumption can be generalized to a time dependent version at the cost of complexity and a potential loss in universality. Also, apart from the estimation method given in Kirchner (2017) and Kirchner and Bercher (2018), there are other estimation algorithms, such as expectation–maximization algorithm (Veen and Schoenberg, 2008), and least squares (Hansen et al., 2015). Since our primary goal in this paper is to arrive at a universal modeling option across different datasets, we choose an exponential triggering function, but one may consider different kernel functions and apply other estimation methods to each individual dataset to explain specific network generating mechanisms.

It is also worthwhile noting that the observed growth patterns serve as motivation to develop variants of the traditional preferential attachment model that use more realistic edge creation assumptions. For example, the common Poisson stable phase in the network growth observed in geographically concentrated networks discussed above, suggests a directed preferential attachment model with Poisson edge increase as proposed in Wang and Resnick (2020). In this model, at each step, a Poisson number of new nodes and new edges are added into the network. By holding the network configuration constant over the creation at each step, this variant is able to capture the growth of networks with timestamps that are only coarsely observed.

4.2 Future Direction

This study gives rise to open questions for future research.

- How can we model the three-phase network evolution theoretically? Interactions among individual users in a network are sometimes modeled by a multivariate Hawkes process, where
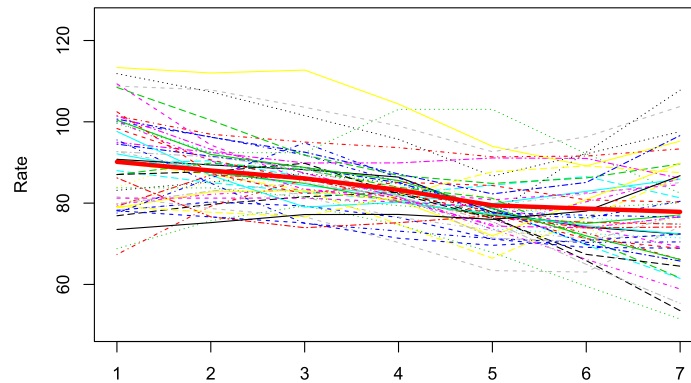
Figure 16: MLE's of daily Poisson rates for the German Wiki data in 2012. Estimates are grouped by week so 1–7 labels on the x-axis represent Monday–Sunday in a week, respectively. The red solid line marks the average rates on each day of the week. We see a decreasing trend over the week, indicating that users are less active during weekends.

  users are influenced by neighbors, authorities and communities (cf. Zhou et al., 2013; Farajtabar et al., 2014). However, considering the whole network evolution, the transition from SEPP to NHPP does not have a systematic analysis. From simulation results, simply superposing SEPPs does not always give a process that is close to Poisson, and the transition seems to depend on the choice of parameters in the SEPP.

- How can we find a more parsimonious model for regional networks in growing or stable phases? One possible way is to investigate any common trend within a week. For example, Figure 16 collects the MLE's of daily Poisson rates in 2012 based on the German Wiki data and groups the estimates by week. Labels 1–7 on the x-axis correspond to Monday–Sunday in a week, respectively. The red solid line highlights the average rate estimates on each day of the week. A decreasing trend is observed, suggesting that people are less active on the German Wiki talk pages during weekends. Hence, one may consider smoothing the daily Poisson rate estimates by taking such weekly pattern into account.

Supplementary Material

We have collected all relevant codes and datasets, and made them available at https://github.com/tw398/NetworkGrowth. This GitHub repository includes: (1) links to the corresponding dataset on KONECT; (2) cleaned Dutch Wikipedia talk data; (3) all relevant codes for data analyses in the paper; and (4) ACF plots which are not presented in the paper.

Acknowledgement

References

Anderson RM, Anderson B, May RM (1992). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.

Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006). Group formation in large social networks: Membership, growth, and evolution. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06*, 44–54. Association for Computing Machinery, New York, NY, USA.

Bacry E, Mastromatteo I, Muzy JF (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01): 1550005.

Bollobás B, Borgs C, Chayes J, Riordan O (2003). Directed scale-free graphs. In: *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Baltimore. 2003, 132–139. New York. ACM.

Daley DJ, Vere-Jones D (2003). *An Introduction to the Theory of Point Processes. Vol. I Probability and its Applications (New York)*. Springer-Verlag, New York, second edition. Elementary theory and methods.

Embrechts P, Liniger T, Lin L (2011). Multivariate Hawkes processes: An Application to financial data. *Journal of Applied Probability*, 48(A): 367–378.

Farajtabar M, Du N, Rodriguez MG, Valera I, Zha H, Song L (2014). Shaping social activity by incentivizing users. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ (Eds.), *Advances in Neural Information Processing Systems*, 2474–2482.

Goh KI, Barabási AL (2008). Burstiness and memory in complex systems. *Europhysics Letters*, 81(4): 48002.

Hansen NR, Reynaud-Bouret P, Rivoirard V (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1): 83–143.

Hawkes AG (2018). Hawkes processes and their applications to finance: A review. *Quantitative Finance*, 18(2): 193–198.

Hawkes AG, Oakes D (1974). A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3): 493–503.

Jo HH, Perotti JI, Kaski K, Kertész J (2015). Correlated bursts and the role of memory range. *Physical Review E*, 92(2): 022814.

Kirchner M (2016). Hawkes and INAR (∞) processes. *Stochastic Processes and Their Applications*, 126(8): 2494–2525.

Kirchner M (2017). An estimation procedure for the Hawkes process. *Quantitative Finance*, 17(4): 571–595.

Kirchner M, Bercher A (2018). A nonparametric estimation procedure for the Hawkes process: Comparison with maximum likelihood estimation. *Journal of Statistical Computation and Simulation*, 88(6): 1106–1116.

Kobayashi R, Tideh RL (2016). Time-dependent Hawkes process for predicting retweet dynamics. In: *Tenth International AAAI Conference on Web and Social Media*.

Krapivsky P, Rodgers G, Redner S (2001). Degree distributions of growing networks. *Physical Review Letters*, 86: 5401–5404.

Krapivsky PL, Redner S (2001). Organization of growing random networks. *Physical Review E*, 63(6): 066123. 1–14.

Leskovec J, Kleinberg J, Faloutsos C (2007). Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1): 2–es.

Mahajan V, Muller E, Bass FM (1990). New product diffusion models in marketing: A review and directions for research. *Journal of Marketing*, 54(1): 1–26.

Martin T, Hofman JM, Sharma A, Anderson A, Watts DJ (2016). Exploring limits to prediction in complex social systems. In: *Proceedings of the 25th International Conference on World Wide Web*, 683–694. International World Wide Web Conferences Steering Committee.

Mishra S, Rizoiu MA, Xie L (2016). Feature driven and point process approaches for popularity prediction. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 1069–1078. ACM.

Mislove A, Koppula HS, Gummadi KP, Druschel P, Bhattacharjee B (2008). Growth of the Flickr social network. In: *Proceedings of the First Workshop on Online Social Networks*, 25–30. ACM.

Mohler G (2013). Modeling and estimation of multi-source clustering in crime and security data. *Annals of Applied Statistics*, 7(3): 1525–1539.

Ogata Y (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401): 9–27.

Porter MD, White G (2012). Self-exciting hurdle models for terrorist activity. *Annals of Applied Statistics*, 6(1): 106–124.

Rizoiu MA, Lee Y, Mishra S, Xie L (2017a). A tutorial on Hawkes processes for events in social media. arXiv preprint: https://arxiv.org/abs/1708.06401.

Rizoiu MA, Xie L, Sanner S, Cebrian M, Yu H, Van Hentenryck P (2017b). Expecting to be hip: Hawkes intensity processes for social media popularity. In: *Proceedings of the 26th International Conference on World Wide Web*, 735–744. International World Wide Web Conferences Steering Committee.

Srijith PK, Lukasik M, Bontcheva K, Cohn T (2017). Longitudinal modeling of social media with Hawkes process based on users and networks. In: Diesner J, Ferrari E, Xu G (Eds.), *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, 195–202. ACM.

Sun J, Kunegis J (2016). Wiki-talk datasets. Universität Koblenz Landau.

Veen A, Schoenberg FP (2008). Estimation of space-time branching process models in seismology using an EM-type algorithm. *Journal of the American Statistical Association*, 103(482): 614–624.

Viswanath B, Mislove A, Cha M, Gummadi KP (2009). On the evolution of user interaction in Facebook. In: *Proceedings of the 2nd ACM Workshop on Online Social Networks*, 37–42. ACM.

Wan P, Wang T, Davis RA, Resnick SI (2017). Fitting the linear preferential attachment model. *Electronic Journal of Statistics*, 11(2): 3738–3780.

Wang T, Resnick SI. A directed preferential attachment model with Poisson measurement. arXiv preprint: https://arxiv.org/abs/2008.07005.

Dutch Wikipedia (2019). https://en.wikipedia.org/wiki/Dutch_Wikipedia. Accessed: 2019-10-27.

German Wikipedia (2019). https://en.wikipedia.org/wiki/German_Wikipedia. Accessed: 2019-10-27.

Wikipedia, Dutch language – Wikipedia, the free encyclopedia (2019). https://en.wikipedia.org/wiki/Dutch_language Accessed: 2019-10-21.

Wikipedia (2019). German language – Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/German_language. Accessed: 2019-10-21.

Zang C, Cui P, Faloutsos C (2016). Beyond sigmoids: The NetTide model for social network

growth, and its applications. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015–2024. ACM.

Zang C, Cui P, Faloutsos C, Zhu W (2017). Long short memory process: Modeling growth dynamics of microscopic social connectivity. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 565–574. ACM.

Zhao Q, Erdogdu MA, He HY, Rajaraman A, Seismic JL (2015). A self-exciting point process model for predicting tweet popularity. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1513–1522. ACM.

Zhou K, Zha H, Song L (2013). Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In: Carvalho CM, Ravikumar P (Eds.), *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, 641–649.