Stockholms universitet

An analysis of the dynamics of a social network

Lei Sun

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2009:1
Matematisk statistik
Oktober 2009

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Master Thesis **2009:1**
http://www.math.su.se

An analysis of the dynamics of a social network

Lei Sun*

Oktober 2009

Abstract

The internet community pussokram.com can be considered as a social network, which developed in time. Data on times when contacts between members took or received are available. We will use two process models to study the dynamic properties of the network. A pure birth process model is used to describe the time between contacts. A probability model for the order, which new contacts are established is also applied. In the analysis, ML-estimates and profile log-likelihood confidence intervals are calculated.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: lesu4833@student.su.se . Supervisor: Åke Svensson.

Abstract

The internet community pussokram.com can be considered as a social network, which developed in time. Data on times when contacts between members took or received are available. We will use two process models to study the dynamic properties of the network. A pure birth process model is used to describe the time between contacts. A probability model for the order, which new contacts are established is also applied. In the analysis, ML-estimates and profile log-likelihood confidence intervals are calculated.

Internetsamfundet pussokram.com kan betraktas som ett socialt nätverk, som utvecklas i tiden. Uppgifter på när deltagare tog eller tog emot besök är tillgängliga för analys. Vi använder två modeller för att studera nätverkets dynamik. En ren födelseprocess används för att beskriva tiden mellan de kontakter en enskild medlem tar. En sannolikhetsmodell för ordningen som nya kontakter tas formuleras och studeras. Analyserna bygger på ML-skattningar och profil log-likelihooder.

Contents

Acknowledgments

I would like to express my deep and sincere gratitude to my supervisor, Professor Åke Svensson at Stockholm University, Department of Mathematic Statistics for his guidance and support. His wide knowledge and abstract way of thinking together with his encouraging and understanding have been of greatest value for me in order to complete the present thesis.

I am deeply grateful to Docent Fredrik Liljeros at Stockholm University, Department of Sociology, for his support throughout this work. His presence has been essential for me.

I owe my loving thanks to my family. Thank you, my parents and boyfriend. Without your encouragement and understanding it would have been impossible for me to finish this thesis.

Shanghai, P.R.China, 2009-10-13

Lei Sun

Chapter 1

Introduction

In this paper, we will use the contact data from the internet community pussokram.com to study the dynamic development of a network of social contacts. The events in this community can be represented as a dynamic network where the members are nodes and the contacts, or visits, are edges. Our aim is to study how the number of visits grows in time. Our goal is to understand the dynamic of this social network. We will use two different approaches. In the first the number of contacts are considered as pure birth processes. For each of n users, $N_1(\cdot), N_2(\cdot), \ldots, N_n(\cdot)$ counts the number of visits. Individual processes started at the time the first visit happened. In the second approach we investigate how the probability to receive or make a new visit depends on the number of previous visits.

In the second chapter, we introduce the background of the internet community pussokram.com and describe the concept of preferential attachment process. This kind of process which is based on the development of degree numbers can be seen as a pure birth process. In the third chapter, we will estimate the jump intensities. And in chapter four we consider a model for how previous visits influence the probabilities for future visits.

To achieve our goal in this paper, we will study two models:

- a pure birth process model which describes the successive times between contacts;
- a probability model where the aim is to describe how a new contact is chosen or from whom a contact is taken. The models try to describe how the choices depend on the number of previous contacts.

Chapter 2

Background and Description

2.1 The Internet Community Pussokram.com

Pussokram.com was a Swedish Internet community for adolescents and young adults to upload social video clip, write blog and communicate with others. Now it exists no more and a new social website, 24suprme.com replaces it. During the spring and summer 2002, the community had around 30 000 active users. The mean user age was 21 years and approximately 70% of the users were female. Both age and sex were self-reported. It is possible to have multiple accounts on the community. A crude check on the number of accounts linked to every unique e-mail address indicates that this was not very common (More than 99.7% of the membership accounts are associated with a unique e-mail address and no e-mail address is associated with more than five accounts) [4].

pussokram.com had a pronounced romantic profile, where

1. Users were encouraged to send messages to others that they were secretly in love with; The provider answered questions related to love and sex posed by the users under the pseudonym Dr.love;

2. The design of the HTML pages made use of a romantic iconography well known to the targeted users (with Valentine's heart, deep red colors, etc. see Figure 2.1). Nevertheless, a quick glance through some of the public guest books revealed that many of the contacts taken were also non-romantic.

There were four major modes of communication at pussokram.com. A brief description of the four types of contacts follows:

1. The *Messages* were in effect intra-community e-mails. Messages were private in the sense that no one in the community, except the sender and receiver, could access

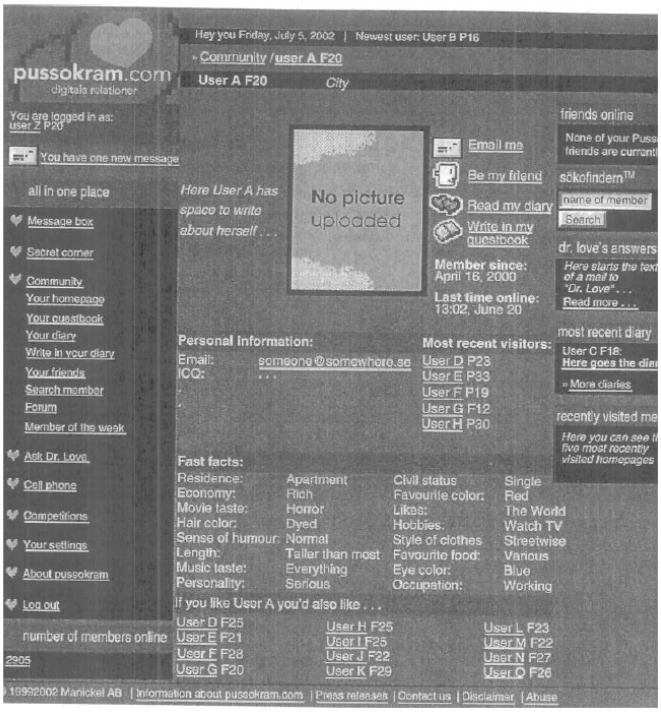

Figure 2.1: Screenshot of a typical user homepage at pussokram.com. "User A", "User B", etc. symbolize user names.

them. Not even information on how many messages other users had received are retrievable for other users.

2. In *Guest book* signing, each user had a guest book that every community member was free to write in.

3. *Flirt* or 'friendship request': user A could ask user B to be her friend. If user B accepts user A's request then they could both easily see if the other is online whenever they are logged into pussokram.com. Information on the friends of a specific user is private to the user only.

4. A *Friendship* relation was established after acceptance of a friendship request, as described above. The friendship network was thus bi-directional. A friendship could be canceled by any of the friends [4].

We will only consider the *Guest book* generated by any of these users. Our data consist of all the users activities on pussokram.com logged for 514 days from 13:39:25h on 13 February 2001 (t=0) to 1:39:25h on 11 July 2002. The smallest time-unit on the log is 1s. The observations can be presented as a list telling who contacted whom at which time. In the Appendix, we listed the first 10 visits and the last 11 visits during the period to show the relationships. This also shows the form of the data that we analyze. We analyze the activities of all users registered at time $t = 0$, as well as the activities of any new users during this time span. Only the activities on the community will be studied; nevertheless this recruitment might induce higher initial growth of active users.

2.2 Statistical Background

In this paper, we will study the dynamic network described above. The individuals who wrote in others guestbooks are active users and the individuals whose guestbooks were written are passive users. One individual can be active user and passive user at the same time.

A preferential attachment process can be seen as a stochastic urn process, meaning a process in which discrete units of wealth, usually called "balls", are added in a random or partly random fashion to a set of objects or containers, usually called "urns" [5]. A preferential attachment process is an urn process in which additional balls are added continuously to the system and are distributed among the urns as an increasing function of the number of balls the urns already have. In the most commonly cases, the number of urns also increases continuously, although this is not a necessary condition for preferential attachment [6]. In this paper, the individuals are the "balls" and they join randomly to

different states (the amounts of contacts), which are "urns" in preferential attachment process. Additional individuals are continuously coming into the system and the states are increasing over time. We will study whether the probability that a user obtains a new contact will be proportional to the previous contact number.

In connection with this dynamic network, some models are formulated and some algorithms are developed for calculation. The analyses are build on the development of degree numbers (the states) and can be described with a pure birth process (A pure birth process is a birth-death process with null death rates for all of the states).

2.3 Description

In the data, id1 and id2 denote the active users respective passive users (cf the Appendix). Their activities are listed according to chronological order. This builds a dynamic network. The users are the nodes in the network. Whenever one contact occured, one link was formed and casted from active user to passive user. We describe a simple network in Figure 2.2. When the participant, id 34215, visited others, an arrow was casted from her to the passive user precisely at the time the contact occured and when she was visited, an arrow was from the active user to her at the time the contact occured. The time is marked over the arrows. When the other participants visited or were visited at certain time, a new arrow would be formed and casted from or to her. Therefore, a dynamic network was formed. We notice that the user, id 34215, visited also herself. Overview the data of *Guest book* for 514 days, the frequencies for different users to make and receive visits are various.

Figure 2.3 shows the number of visits made and received. Each point corresponds to a specific user. The horizontal coordinates are the number of visits made and the vertical coordinates are the number of visits received.

The points near the horizontal axis (e.g. in the ring on horizontal axis) indicate that the corresponding users made visits frequently but received rarely. The situation is the opposite for the users near vertical axis (e.g. in the ring on vertical axis). The users were visited fairly many times, however they rarely visited others. The overall situation for majority users is that less visits they received than made.

There are 18063 active users during the 514 days. And for 92.04%, the total number of visits is less than 10. In the other words, a majority of active users did not make visits frequently. Moreover, some active users visited different members at the same day. It means that the development of the visit numbers is kind of leap-style instead of step by step.

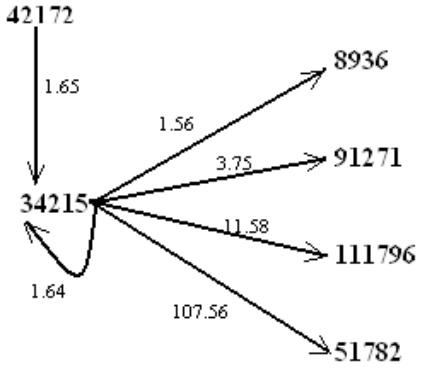

Figure 2.2: The simple dynamic network of user 34215

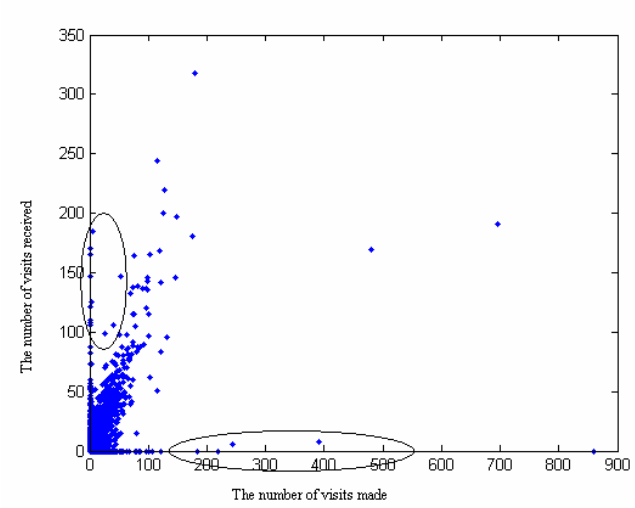

Figure 2.3: The number of visits made and the number of visits received

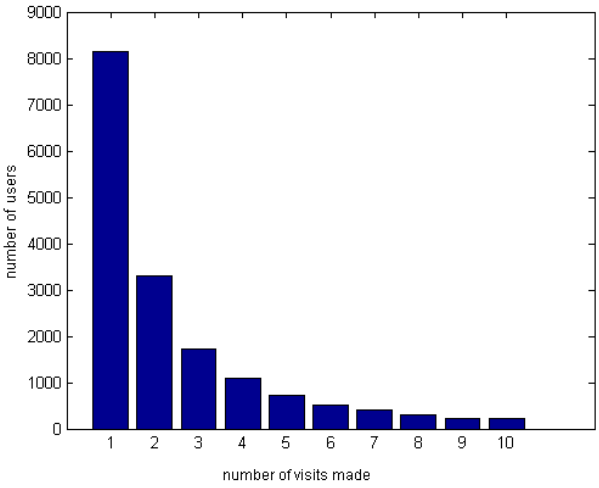

Figure 2.4: The number of users with different number of visits made

The histogram in Figure 2.4 reports the number of users with 1 through 10 visits. The horizontal coordinates are the number of visits made and the vertical coordinates are the number of active users. State i gives that the users made i visit(s) during the period. The proportion is decreasing with the increase of states. Many users generated only one visit during the 514 days.

There are 13774 passive users. Not all of them were active users. 88.34% passive users did not receive visits frequently, less than 10 during the period. Some passive users were visited by several users at the same time.

The histogram in Figure 2.5 shows the number of users with 1 through 10 visits. The horizontal coordinates are the number of visits received and the vertical coordinates are the number of passive users. The proportion of users decreases with the increase of states. The amounts of active users as well as passive users are descreasing.

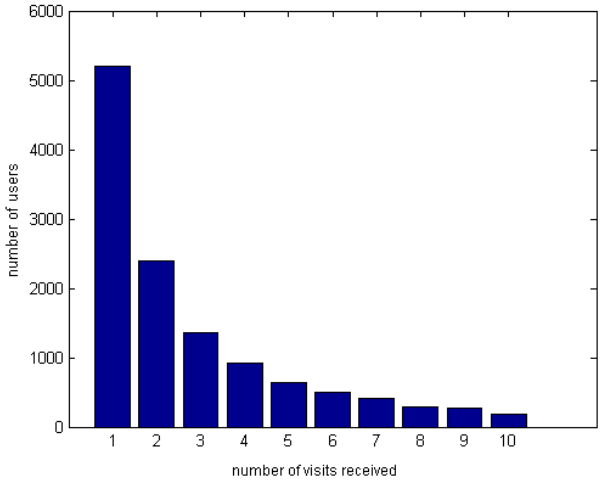

Figure 2.5: The number of users with different number of visits received

Chapter 3

Pure Birth Process Models

The analyses in this chapter are based on the assumption that the users stayed at state i for a random time that is exponentially distributed. λ_i is defined as the intensity to leave state i and $1/\lambda_i$ is the mean time to stay. A particular feature is that a participant can make serveral contacts at the same time. This implies that, opposite what is commonly assumed in birth process, there are jumps from state i to $i+k$, where $k > 1$.

3.1 A Non-parametric Model for the Jump Intensities

For the convenience of analysis, we estimate only the first 20 visiting intensities without any assumption of a parametric representation of the intensities. The intensities are estimated by the ratio of the amount of users jumping from state i and the time of users staying at state i. The estimated visit intensities shown i Table 3.1 are ML-estimated (cf the likelihood equation 3.3 below). As we mentioned before, this kind of jump is leap-style. If the user jumped from state i to $i+k$, she would not be at state $i+1, ...i+k-1$ and not spend any time in these states either.

$$\lambda_i = \frac{\sum \text{jump from state } i}{\sum \text{time at state } i} \tag{3.1}$$

The results are shown in Table 3.1.

A crude check on the intensities indicates that the intensities are increasing with the number of previous contacts which means that the users will stay at a higher state for a shorter time.

Table 3.1: The intensities to make visits at the first 20 states

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
0.0034	0.0060	0.0085	0.0106	0.0129	0.0152	0.0170	0.0183	0.0216	0.0206
λ_{11}	λ_{12}	λ_{13}	λ_{14}	λ_{15}	λ_{16}	λ_{17}	λ_{18}	λ_{19}	λ_{20}
0.0215	0.0285	0.0293	0.0292	0.0261	0.0336	0.0311	0.0291	0.0290	0.0443

The ML-estimates of the visited intensities for the users are shown in Table 3.2.

Table 3.2: The intensities to receive visits at the first 20 states

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
0.0489	0.0062	0.0087	0.0105	0.0121	0.0140	0.0147	0.0178	0.0166	0.0218
λ_{11}	λ_{12}	λ_{13}	λ_{14}	λ_{15}	λ_{16}	λ_{17}	λ_{18}	λ_{19}	λ_{20}
0.0207	0.0251	0.0264	0.0268	0.0297	0.0331	0.0311	0.0312	0.0317	0.0423

Obviously, the first intensity is much larger than others. This indicates that the users stayed at the first state for a very short time before jumping to the next state. From the second state, the jump intensities are increasing with the increase of the states, which indicates that the higher the state, the shorter time the users will stay.

3.2 Likelihood-Equations

We have estimated the jump intensities without any model and notice that the intensities develop over the states. Now we try to find a regular pattern of them.

Assume that user j starts a birth process at the time of the first visit and denote the jump intensities with λ_1, λ_2, ... Let $\tau_{j,i}$ be the time the user j spends at the state i, $i = 1, 2, ..., N_j$. N_j is the last state user j stays. Then

$$f(\tau_{j,i}) = \lambda_i e^{-\lambda_i \tau_{j,i}}, \tag{3.2}$$

$\tau_{j,i}$, $i = 1, 2, \ldots, N_j$ are independent and exponentially distributed random variables. This property is used to derive the likelihood. The likelihood function for user j is then

$$L_j = \prod_{i=1}^{N_j - 1} (\lambda_i e^{\lambda_i \tau_{j,i}})^{Z_{j,i}} exp(-\lambda_{N_j} \tau_{j,N_j}), \tag{3.3}$$

where

$$Z_{j,i} = \begin{cases} 1 & \text{if user } j \text{ jumped from state } i \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

The log-likelihood for user j is

$$l_j = ln(L_j) = ln(\prod_{i=1}^{N_j-1} (\lambda_i e^{-\lambda_i \tau_{j,i}})^{Z_{j,i}} exp(-\lambda_{N_j} \tau_{j,N_j})) \tag{3.5}$$

and the log-likelihood for all of the users is

$$\begin{aligned} l = \sum_j l_j = \sum_j \sum_{i=1}^{N_j-1} Z_{j,i} ln(\lambda_i) - \sum_j \sum_{i=1}^{N_j} Z_{j,i} \lambda_i \tau_{j,i} = \\ = \sum_{i=1}^{N-1} Z_{i.} ln(\lambda_i) - \sum_{i=1}^{N} Z_{i.} \lambda_i \tau_{i.} = \sum_{i=1}^{N-1} Z_{i.} ln(\lambda_i) - \sum_{i=1}^{N} \lambda_i T_{i.} \end{aligned} \tag{3.6}$$

In Equation 3.6, $Z_{i.}$ is the number of jumps from state i; $T_{i.}$ is the total time all users spending at the state i. If some users jumped over state i, the time for the users at state i is null. N is the largest number of visits among all of the users and equal to 859 which means that some user made 859 visits during the 514 days.

3.3 A Parametric Model for the Jump Intensities

Assume the jump intensities model is λ_i=γi^δ as is done in [3], and insert it into 3.6 in the previous section. Then

$$l = \sum_{i=1}^{N-1} Z_{i.} ln(\gamma i^\delta) - \sum_{i=1}^{N} \gamma i^\delta T_{i.} \tag{3.7}$$

Differentiating Equation 3.7 with respect to γ respective δ yields

$$\frac{\partial l}{\partial \gamma} = \frac{1}{\gamma} \sum_{i=1}^{N-1} Z_{i.} - \sum_{i=1}^{N} i^\delta T_{i.} \tag{3.8}$$

and

$$\frac{\partial l}{\partial \delta} = \sum_{i=1}^{N-1} Z_{i.} ln(i) - \gamma \sum_{i=1}^{N} i^\delta T_{i.} ln(i). \tag{3.9}$$

Equating (3.8) to 0 gives the solution of γ which maximizes the log-likelihood function of γ. That is

$$\gamma = \frac{\sum_i^{N-1} Z_{i.}}{\sum_i^{N} i^\delta T_{i.}}. \tag{3.10}$$

Insert the solution (Equation 3.10) into (3.7) and yields a equation with δ

$$l(\delta) = \sum_{i}^{N-1} Z_{i.}(ln(\sum_{i}^{N-1} Z_{i.}) + \delta ln(i) - ln(\sum_{i}^{N} i^{\delta} T_{i.})) - \sum_{i}^{N-1} Z_{i.}. \tag{3.11}$$

Defferentiating the above function with respect to δ yields

$$\frac{\partial l}{\partial \delta} = \sum_{i}^{N-1} Z_{i.} ln(i) - \sum_{i}^{N-1} Z_{i.} \frac{\sum_{i}^{N} i^{\delta} T_{i.} ln(i)}{\sum_{i}^{N} i^{\delta} T_{i.}} \tag{3.12}$$

Equating the above function to 0 gives $\hat{\delta} = 0.7023$, and insert it into Equation 3.10 yields $\hat{\gamma} = 0.0048$. Table 3.3 shows the first 20 visiting intensities estimated by this model.

Table 3.3: The intensities to make visits at the first 20 states

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
0.0048	0.0077	0.0103	0.0126	0.0147	0.0168	0.0187	0.0205	0.0223	0.0204
λ_{11}	λ_{12}	λ_{13}	λ_{14}	λ_{15}	λ_{16}	λ_{17}	λ_{18}	λ_{19}	λ_{20}
0.0256	0.0273	0.0288	0.0304	0.0319	0.0334	0.0348	0.0362	0.0376	0.0390

Comparing Table 3.1 and Table 3.3, the numerical in these two tables are very similar. The following figure shows their tiny differences.

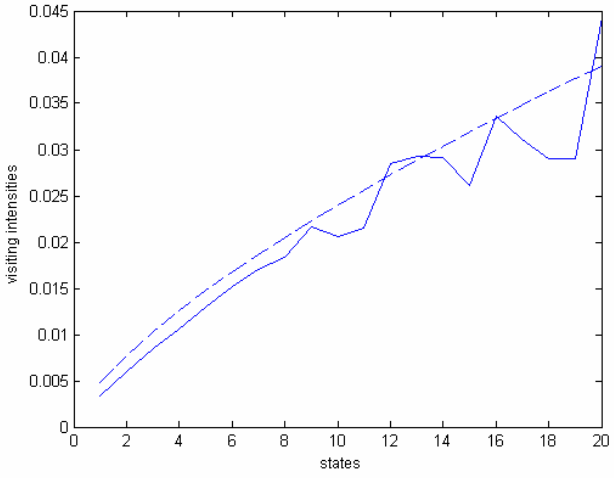

Figure 3.1: The differences between jump intensities estimated using the non-parametrical approach (solid line) and the parametric approach (dotted line)

In Figure 3.1, the horizontal coordinates are the states and the vertical coordinates are the visit intensities. The dotted line is the jump intensities estimated by the parametric model and the solid line is by non-parametric model.

The dotted line lies above the solid line for most states indicates that the visit intensities estimated by parametric model are a little bit larger than the intensities without model. The dotted line is nearly linear, meaning that the time users stay at states decreases almost proportionally with the increase of states.

With the same model, we estimate the intensities to receive visits and yield $\hat{\delta} = 0.6410$ and $\hat{\gamma} = 0.0051$. The intensities of the first 20 states are shown in Table 3.4.

Table 3.4: The intensities to be visited at the first 20 states

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
0.0051	0.0079	0.0103	0.0124	0.0143	0.0161	0.0177	0.0193	0.0208	0.0223
λ_{11}	λ_{12}	λ_{13}	λ_{14}	λ_{15}	λ_{16}	λ_{17}	λ_{18}	λ_{19}	λ_{20}
0.0237	0.0250	0.0264	0.0276	0.0289	0.0301	0.0313	0.0325	0.0336	0.0347

Comparing the visited intensities with two estimations in Figure 3.1, the horizontal coordinates in the figure are the states and the vertical coordinates are the visited intensities. The dotted line is the jump intensities estimated by parametric model and the solid line is by non-parametric model.

The largest difference between the two lines is that the the first state has the largest jump intensities among the 20 states in the solid line. The solide curve falls down from the first states to the second state and then increases. Meanwhile, the dotted line is nearly linear and increase all the way. The reason to the difference is probably that the time between the first two visits is short in reality. So the visited intensity is estimated fairly large with non-parametric model.

It seems that the first state is a special situation. In that case we adjust the model so that it is only valid from state 2. $\hat{\delta}$ decreases to 0.5908 and $\hat{\gamma}$ increases to 0.0061. The new comparing figure is shown below.

In Figure 3.3 the dotted line is the visited intensities estimated by parametric model and the solid irregular curve is by non-parametric model. The two curves are very near each other. But for most states, the visited intensities estimated by parametric model are a little larger than the ones by non-parametic model.

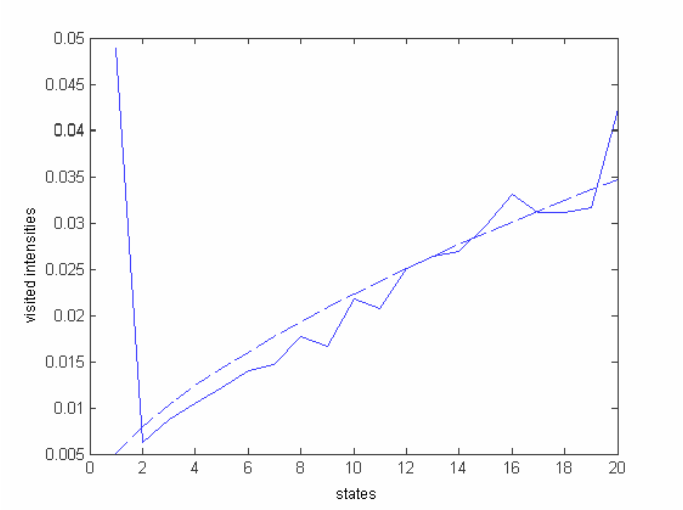

Figure 3.2: The differences between jump intensities estimated using the non-parametrical approach (solid line) and the parametric approach (dotted line)

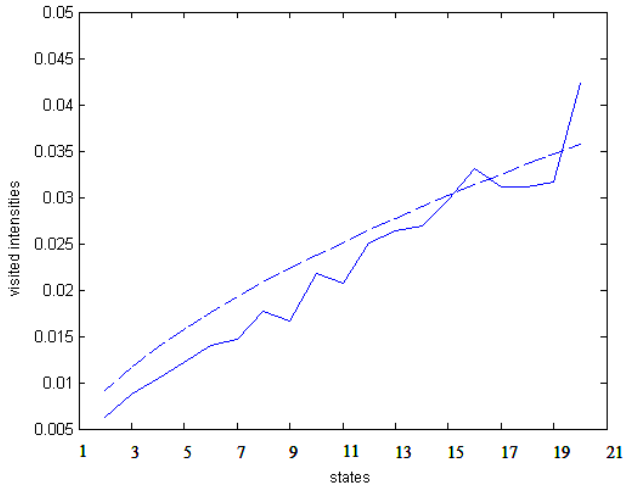

Figure 3.3: The differences between jump intensities estimated using the non-parametrical approach (solid line) and the parametric approach (dotted line) without state 1

3.4 Confidence Intervals Based on Profile Likelihoods

We have estimated the value of $\hat{\delta}$ in the previous section. However, in practice it is more informative to construct confidence interval for parameters than to test hypothesis about their values. We illustrate Equation 3.11 with profile log-likelihood interval.

The profile log-likelihood function $l(\delta)$ is the function of δ that gives the maximal likelihood obtainable for each δ (the maximum taken over the parameter δ) [2].

Evaluated at $\delta = \hat{\delta}$, the profile log-likelihood confidence interval for δ is the set of δ for which

$$-2[l(\delta) - l(\hat{\delta})] \leq \chi_1^2(95\%)[1]. \tag{3.13}$$

The differences between $l(\delta)$ and $l(\hat{\delta})$ are shown in the following figures.

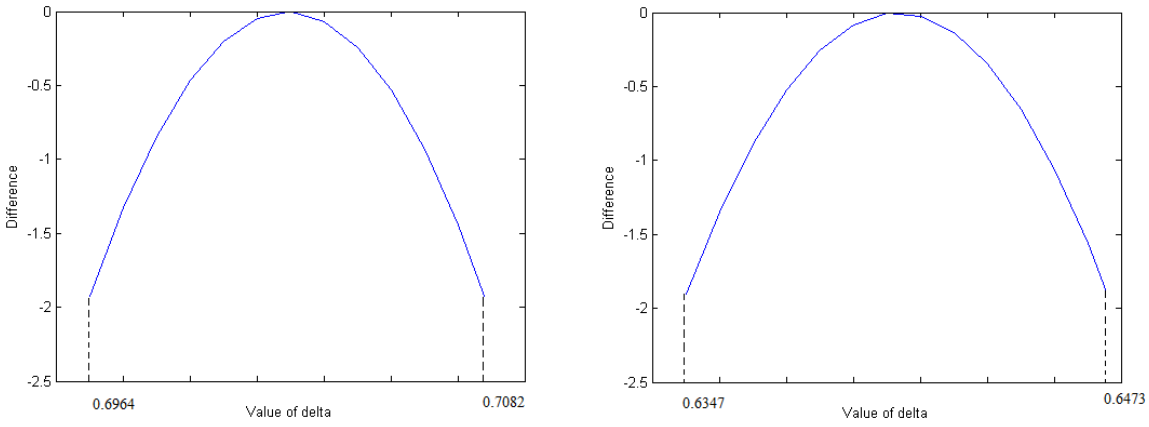

Figure 3.4: The profile confidence interval of visit respective visited log-likelihood function

In the two figures above, the smallest numerical-unit on the horizotal coordinate is 0.001. A 95% confidence interval is formed by all δ such that $[l(\delta) - l(\hat{\delta})] \geq -\frac{1}{2}\chi_1^2(95\%)$. The figure on the left is for the visit intensities and figure on the right is for the visited intensities. The vertexes of the two curves indicate where δ is exactly equal to $\hat{\delta}$ and the log-likelihood functions are maximized. In that case, $\hat{\delta}$ is equal to 0.7023 respective 0.6410 for the visit and visited intensities.

Chapter 4

Probability Model for the Order

We have analyzed the pure birth process model for jump intensities from state i to state $i + k$. The jump intensities increase almost proportionally with the increase of state i. The higher state, the shorter time users will stay. It seems that the higher state one participant stays, the more eager she is to jump. We will now consider an analysis which only takes the order in which new connections are established into account.

4.1 Probability to Receive Visits

In this section, we will study the probability to receive visits at different states.

Assume that the probability for user i to receive visits depends on the amount of visits has been made until time t. We assume that

$$p_{i,t} = \frac{b_{i,t}^{\alpha}}{\sum b_{k,t}^{\alpha}}, \tag{4.1}$$

where $b_{i,t}$ is the visited number of a specific user i in the system until time t and $b_{k,t}$ is visited number of user k ($k = 1, 2, \dots$). α is a parameter to describe how the probability varies depending on the visits already received.

If $\alpha = 0$, the probability to receive a new visit is equal for each user. If not, the probability varies over time because $b_{k,t}$ is not a constant, varies over time and depends on the number of visits a person has made previously. To simplify the computation, the time is limited to be integral days, $t = 1, 2, \dots, 514$.

Assume that users receiving visits do not depend on each other. So the probability for different users to receive a new visit is independent. Because of it, the likelihood function of the probability for a specific user i is

$$L_i(\alpha) = \prod_t \frac{b_{i,t}^{\alpha}}{\sum_k b_{k,t}^{\alpha}}, \tag{4.2}$$

then the log-likelihood function is

$$l_i(\alpha) = \sum_t \alpha ln(b_{i,t}) - \sum_t ln(\sum_k b_{k,t}^{\alpha}) \tag{4.3}$$

.

For all the users in the system until time t, the log-likelihood function is the sum of log-likelihood funtions of every user

$$l(\alpha) = \sum\nolimits' (\sum_t \alpha ln(b_{i,t})) - \sum\nolimits' (\sum_t ln(\sum_k b_{k,t}^{\alpha})), \tag{4.4}$$

where $\sum'$ indicates the summation over the indices of the users that received visits until time t.

Differentiating Equation 4.4 with respect to α yields

$$\frac{\partial l}{\partial \alpha} = \sum\nolimits' \sum_t ln(b_{i,t}) - \sum\nolimits' \sum_t \frac{\sum_k b_{k,t}^{\alpha} ln(b_{k,t})}{\sum_k b_{k,t}^{\alpha}} \tag{4.5}$$

Since the dynamic graph that we are studying is very large. It has not, for numerical reasons, been possible to carry through these calculations. We have here only made a simple analysis on the contacts taken during the last day by solving the ML-equation

$$\frac{\partial l}{\partial \alpha} = \sum\nolimits' (ln(b_{i,514})) - H \frac{\sum_k b_{k,514}^{\alpha} ln(b_{k,514})}{\sum_k b_{k,514}^{\alpha}}, \tag{4.6}$$

where $\sum'$ denotes summation over the indices of the persons that are visited at the last day; H is the number of users who received visits at the last day, and $\sum_k$ is over all persons in the system.

Equating the above function to zero yields $\hat{\alpha}$ that maximizes the log-likelihood function (Equation 4.4) based on the contacts during the last day. As we mentioned before, it is more interesting to consider the confidence interval of $\hat{\alpha}$ than its value. If $[l(\alpha) - l(\hat{\alpha})] \geq -\frac{1}{2}\chi_1^2(95\%)$, α belongs to the 95% confidence interval.

The differences between $l(\alpha)$ and $l(\hat{\alpha})$ is shown in Figure 4.1. In the figure, the smallest numerical-unit is 0.0001 on the horizontal axis. The vertical coordinates are the differences between $l(\alpha)$ and $l(\hat{\alpha})$. We know from the figure that the confidence interval of $\hat{\alpha}$ is approximately between -0.1156 and 0.6169. When $\hat{\alpha}$ equals to 0.2690, the log-likelihood

function is maximized. However, the fact that $\hat{\alpha} = 0$ is in the interval indicates that the hypothesis that the probability to receive a new visit is independent of the number of previous visits cannot be rejected.

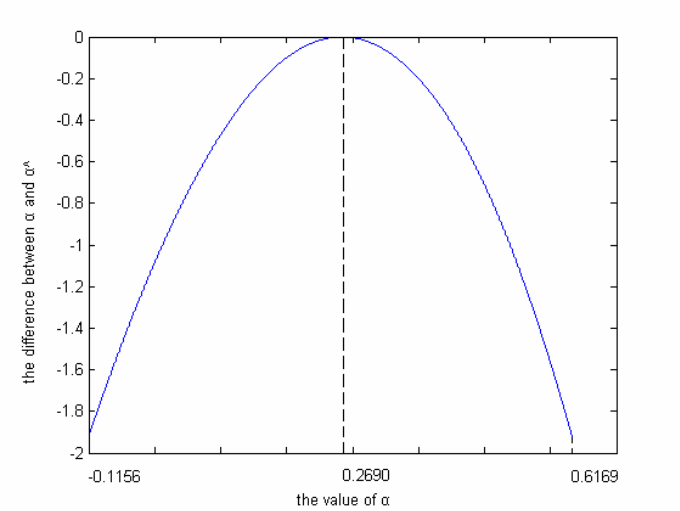

Figure 4.1: The differences between $l(\alpha)$ and $l(\hat{\alpha})$

4.2 Probability to Make Visits

We have illustrated the probability to receive a new visit and come to the conclusion that different passive users have the same probability to receive a new visit. Then how is the situation for users to visit others?

Assume that the probability for user i to visit others until time t is

$$p_{i,t} = \frac{b_{i,t}^{\beta}}{\sum_k b_{k,t}^{\beta} + c^{\beta}}, \tag{4.7}$$

where $b_{k,t}$, $k = 1, 2, \ldots$ is the amount of visits of user k in the system until time t and $b_{i,t}$ is the amount of visits of a specific user i. β is a parameter and c is constant. In this paper, c is assumed to be 1.

Therefore, the likelihood function of user i is

$$L_i(\beta) = \prod_t \frac{b_{i,t}^{\beta}}{\sum_k b_{k,t}^{\beta} + c^{\beta}} \tag{4.8}$$

For all of the users in the system until time t, insert $c = 1$, the log-likelihood function is

$$l(\beta) = \sum\nolimits' \sum_t \beta ln(b_{i,t}) - \sum\nolimits' \sum_t ln(\sum_k b_{k,t}^{\beta} + 1), \tag{4.9}$$

where $\sum'$ indicates the summation over the indices of the users that made visits until time t.

Differentiating Equation 4.9 with respect to β yields

$$\frac{\partial l}{\partial \beta} = \sum\nolimits' \sum_t ln(b_{i,t}) - \sum\nolimits' \sum_t (\frac{\sum_k b_{k,t}^{\beta} ln(b_{k,t})}{\sum_k b_{k,t}^{\beta} + 1}) \tag{4.10}$$

We only make analysis on the visits made at the last day by solving the ML-equation

$$\frac{\partial l}{\partial \beta} = \sum\nolimits' ln(b_{i,514}) - H \frac{\sum_k b_{k,514}^{\beta} ln(b_{k,514})}{\sum_k b_{k,514}^{\beta} + 1} \tag{4.11}$$

where $\sum'$ denotes summation over the indices of the persons that made visits at the last day; H is the number of users who made visits at the last day, and $\sum_k$ is over all persons in the system.

Equating the above function to zero yields $\hat{\beta}$ which maximizes the log-likelihood function based on the contacts during the last day. As we did with $\hat{\alpha}$, we calculate a 95% profile confidence interval of $\hat{\beta}$. Figure 4.2 shows the differences between $l(\beta)$ and $l(\hat{\beta})$.

In Figure 4.2, the vertical coordinates are the differences between $l(\beta)$ and $l(\hat{\beta})$. The minimun numerical-unit on horizontal is 0.0001. The confidence interval of $\hat{\beta}$ is between 0.3835 and 1.0262. When $\hat{\beta} = 0.7193$, Equation 4.9 is maximized.

The fact that $\hat{\beta} = 0$ is not in the 95% profile confidence interval indicates that we can reject the hypothesis that different users have the same probability to make a new visit.

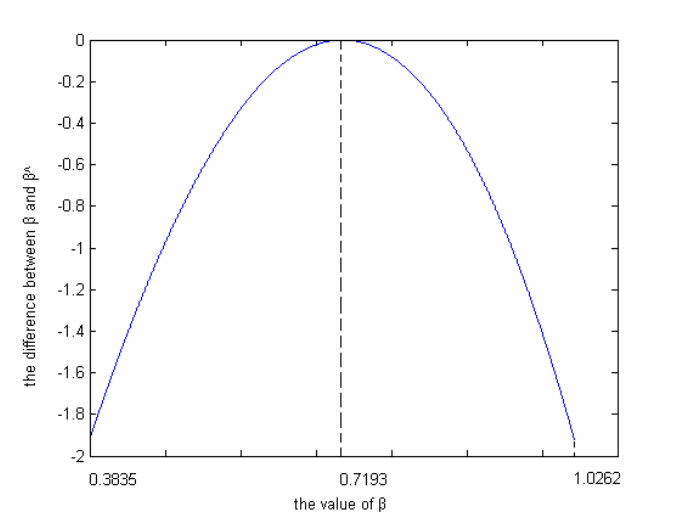

Figure 4.2: The differences between $l(\beta)$ and $l(\hat{\beta})$

Chapter 5

Conclusion and Discussion

With the previous results, we come to the following conclusions.

With the pure birh process model to analyse the time between two visits to make or to receive, we got that the amount of visits has a slightly slower than linear development. With the jump intensities, we can also estimate approximately the time one individual needs to reach a certain state.

With the probability model describing the relationship between taken and received visits, we got different results for the two kinds of probabilities. Because $\hat{\alpha} = 0$ is in the 95% confidence interval of $\hat{\alpha}$, we come to the conclusion that the probability to be visited is the same for all of the users regardless of the number of visits received. This seems natural since the number of visits is not known to other users. However, $\hat{\beta} = 0$ is out of the 95% confidence interval, so the probability for different users to make visit depends on the number of previous visits. The positive values of $\hat{\beta}$ indicates that the more previous visits one individual has, the higher probability for her to make a new visit.

However, in our analysis, there are some defects. First of all, the data between day 495 and day 507 are missing. The website was maybe under maintenance, so users could not log in and did anything. This would make error in our calculation.

Second, the estimates with non-parametric model and parametric model differ much from the fifth state. The intensities estimated by two ways in general matches each other, specially the first four states. However, the estimates without model fluctuate much from the fifth state and over. Our opinion is that from the fifth state, less than 1000 users left in the system, and even less users at the later states (The reducing of the users are shown in Figure 2.4 and Figure 2.5). The shrinkage of data size affects the intensity estimates without model. So the less data, the more inaccurate intensities would be estimated.

The visited intensity estimated with non-parametric model at the first state differs from the estimate with model, although there are more than 5000 users in the system. The reason is probably that in reality users are passionate about visiting others in the beginning, the time interval between two visits is small. Thus the total time for users to be visited at the first state is short and we yield a fairly large jump intensity at the first state.

At the later states, users' enthusiasm waned. Less and less users in the system visited the guest books. The total time at the later accordingly decreases and the intensities increase. The curves of visiting intensities and visited intensities are similar. This identify our estimates from another aspect.

Third, when we illustrate the relationship between two specific users, We assume that the probabilities for users to make visits or to receive visits are independent. This assumption is a little over-theoretical. For example, if two users knew each other, then the probability to visit each other would be probably higher than to visit other non-familiar users. However when we estimate the probabilities, we did not account this kind of factor into our analysis. Therefore there will be a certain gap between the estimates and reality.

The last but not least defect is when we estimate the parameter α and β, we did not take all the days. Because of the limitations of computing capacity, we just calculate the last day and make a simple analysis on the contacts occurs during the last day. There is an error between the result and reality for certain. And unfortunately, we do not know how large the error is. We could yield a more accurate result if we could approach a more effective computing system.

Appendix

Table 5.1: The first 10 visits

id1	id2	time
34215	8936	1.56
34215	34215	1.64
42172	34215	1.65
42183	42172	1.71
8838	8560	1.72
42172	8560	1.72
8560	8560	1.72
123154	8560	1.73
42183	8560	1.74
8936	8936	1.75

Table 5.2: The last 11 visits

id1	id2	time
172430	174138	513.37
143842	174157	513.37
162363	174157	513.37
174233	137070	513.38
174195	174157	513.39
172430	174232	513.39
174233	174157	513.39
28897	151940	513.43
162229	173284	513.46
138829	168547	513.47
172899	173503	513.50

Bibliography

[1] Alan Agresti. *Categorical Data Analysis*. Wiley, 2002.

[2] David Clayton and Michael Hills. *Statistical Models in Epidemiology*. Oxford University Press, 2002.

[3] Birgitte Freiesleben de Blasio, Åke Svensson, and Fredrik Liljeros. Preferential attachment in sexual networks. *Proceedings of the National Academy of Sciences of the United States of America*, (104):10762–10767, 2007.

[4] Petter Holme, Christofer R. Edling, and Fredrik Liljeros. Structure and time evolution of an Internet dating community. *Social Networks*, (26):155–174, 2004.

[5] M.E.J. Newmana. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, (46):323–351, 2005.

[6] G.U. Yule. A Mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London*, (213):21–87, 1925.