4-2012

Modeling Latent Relationships in the myGamma Network

Mauricio Sadinle

Mike Finegold

Stephen E. Fienberg

Ee-Peng Lim
Singapore Management University, eplim@smu.edu.sg

Modeling Latent Relationships in the myGamma Network

Mauricio Sadinle, Carnegie Mellon University
msadinle@andrew.cmu.edu
Mike Finegold, Carnegie Mellon University
mfinegol@andrew.cmu.edu
Stephen E. Fienberg, Carnegie Mellon University
fienberg@andrew.cmu.edu
Ee-Peng Lim, Singapore Management University
eplim@smu.edu.sg

April, 2012

LARC-TR-02-12

LARC Technical Report Series: http://smu.edu.sg/centres/larc/larc-technical-reports-series/

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Carnegie Mellon University

Modeling Latent Relationships in the myGamma Network

Mauricio Sadinle

Faculty members: Mike Finegold and Stephen E. Fienberg

Carnegie Mellon University

External advisor: Ee–Peng Lim

Singapore Management University

Advanced Data Analysis Report
Department of Statistics
Carnegie Mellon University

April 29, 2012

Chapter 1

The myGamma Network

1.1 Introduction

myGamma[1] is a mobile social networking service provided by BuzzCity[2], a company based in Singapore. The myGamma network was started in 2003 mainly in Singapore, Malaysia and Thailand. By August 2010, 4,659,108 users had accessed the network. Among these users, 583,570 joined the network between January and the first week of August 2010 and 62,291 accessed the network during the first week of August 2010. BuzzCity has characterized the users of myGamma as people that access the Internet primarily via mobile phones, living in emerging markets or working in the blue collar sector in wealthier nations[3].

Regarding the structure of myGamma users as a social network lead us to the consideration of several possibilities. First of all, users can add other users to their friend list. In such a case a directed link is formed from the user who claimed the friendship (source) to the user added (target). Later on, this action can be reciprocated and the second user can add the first user to his/her friend list. An user can also block another user, in which case a foe link is formed from the user who blocked (source) to the blocked user (target). Another way to think on the structure of myGamma as a social network is as a valued directed network, where each link between two users is the number of messages that one user sent to another. The structure of the myGamma network has been studied recently by Ee–Peng Lim and his collaborators at Singapore Management University (Leung et al., 2010; On et al., 2010).

We explore some characteristics of the myGamma network, specifically the possible role of countries as communities.

1.2 Distribution of Users by Country

In the left hand side of Figure 1.1 we present a barplot of the 20 countries with the largest number of myGamma users in the period 2003 – August 2010. In Figure 1.2 we present a world map of the number of users during the same period. We can see that myGamma has the most important number of users in India, with almost 1.2 million users during the whole period. Also the Southeast Asian countries have large numbers of users, as well as some African countries such as South Africa, Kenya, Nigeria, Egypt and Libya, among others.

In the right hand side of Figure 1.1 we present the barplot of the 20 countries with the largest rate of users per 1000 people[4]. The high rate of users of Monaco and Brunei suggest that these users probably do not belong to the permanent populations of these countries. In the case of Brunei, it is the tenth country with the largest

[1] http://m.mygamma.com/
[2] http://www.buzzcity.com/
[3] http://www.buzzcity.com/l/coverage/MobileSocialNetworking.pdf
[4] Taking population estimates from http://esa.un.org/UNPP/

number of users and accounts for around 140,000 users. This fact is interesting given that Brunei's population is just around 400,000 people according to a 2010 UN estimate[5]. Such a number of users could reflect the popularity of myGamma among temporally international workers in Brunei, but at this point this is just a hypothesis.

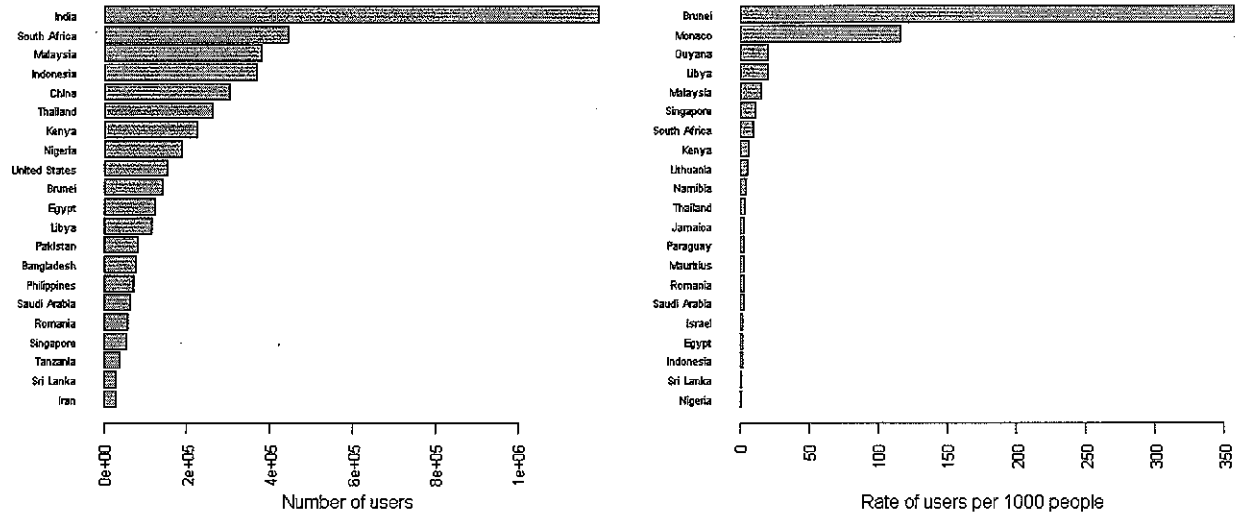


Figure 1.1: Left: Top 20 countries by number of myGamma users, 2003–August 2010. Right: Top 20 countries by rate of myGamma users, 2003–August 2010.

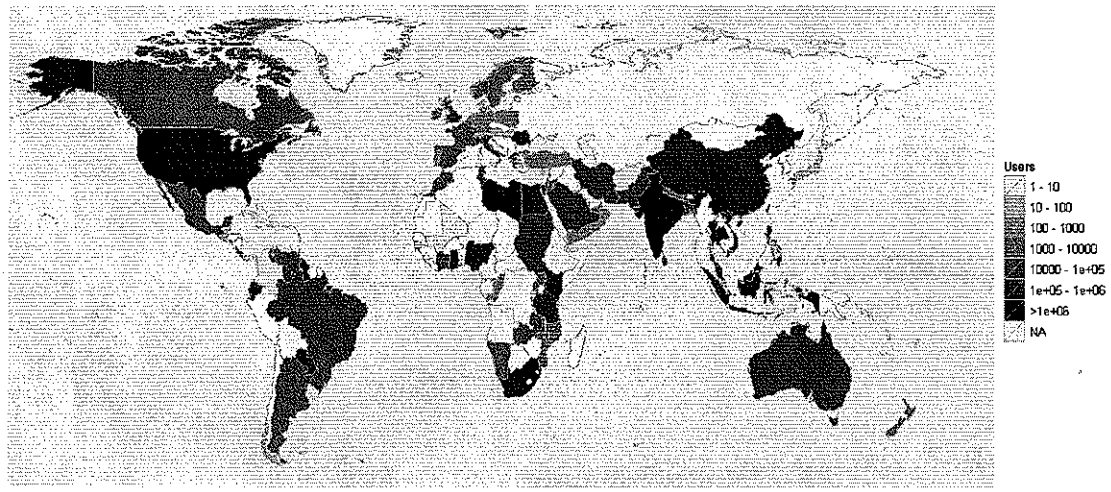


Figure 1.2: myGamma users around the world, 2003–August 2010.

[5]http://esa.un.org/UNPP/

1.3 Worldwide Evolution of the Network

In Figure 1.3 we present the number of users that joined myGamma for each year from 2003 to 2010. As we mentioned above we can see that in the beginning all the myGamma users were mainly located in southeastern Asian countries. From 2003 to 2005 the expansion of the network occurred mainly in southeastern Asian countries, China, India and Australia. In 2006 this pattern changed and myGamma gained an important number of users all around the world. Also, in 2006 South Africa starts to appear with a huge number of users. From 2006 to 2010, broadly speaking, we can see a continuous expansion of the network among African, Arabic countries and the US. Something to notice is that during this period the number of users joining the network has been decreasing yearly for South Africa and China.

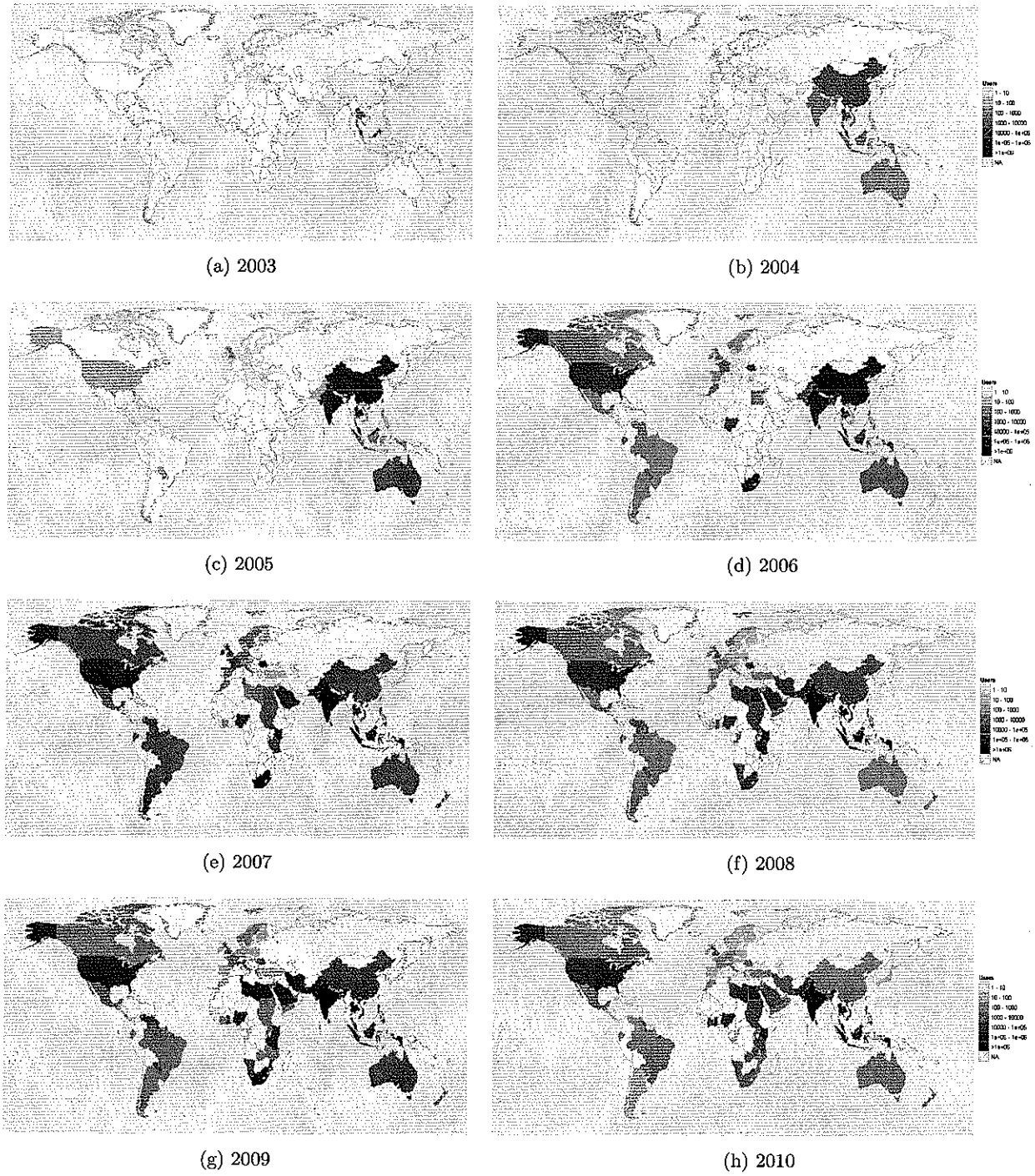
1.4 Source–Target Relations by Country

Since each friend link is formed by a source and a target, it is possible to build a bidimensional contingency table of the links, such that it classifies in one factor the countries of the sources and in the second factor the countries of the targets. In Figure 1.4 we can see a fluctuation plot in which each column represents the distribution of targets for all the links with source in each country. The darkest the box in the intersection of each pair of countries, the largest the proportion of links targeting individuals in the country of the row, among the links with source in the country of the column. For simplicity, the countries of the columns are represented by a number that can be seen in the names of the rows. Figure 1.4 shows the information of the users that accessed myGamma during August 2010 for the countries with more than 20 active users during August 2010. The countries are ordered by the number of users that accessed the network during August 2010, so the countries that appear at the top are the ones with the largest numbers of active users.

We can see from Figure 1.4 that the structure of the myGamma network is far from a country–community structure in which we would expect most of the links from one country to target users in the same country. We can see that for almost all the countries a large proportion of the targets are users in countries where myGamma is popular, i.e. India, Kenya, Thailand and Indonesia. A first hypothesis that we could state is that for many countries, many users of myGamma are migrants from the countries where this network is popular. For instance, almost all the links from users in Morocco (42) do not target users in Morocco, but mostly users in Libya and Egypt. This situation is similar for many other countries. Most of the links from wealthy countries actually target users in India and Kenya and almost do not target users into the same country. A broad list of such countries are US (16), UK (28), Australia (37), France (45) and Canada (47). Also, in Figure 1.4 the countries coded between 33 and 49 have a number of users between 20 and 100. We can see from this figure that for these countries, where myGamma is not very popular, the proportion of links targeting users in countries where myGamma is popular is even higher than for the other countries, which may also be a signal that the users in these countries are immigrants.

1.5 Evolution of Source–Target Relations by Country

In order to continue exploring the possible role of migration as a catalyst of the spread of myGamma, in Figure 1.5 we present the same plot presented in Figure 1.4, but in this case each plot represents the links by the year in which the users joined the network. The color scale and the order of the countries is the same as presented in Figure 1.4, despite we do not present this information in the figure given the constraints of space. We can see that for 2003 the four countries that appear, which are China, Malaysia, Singapore and Thailand, follow almost perfectly a country–community structure. Later in 2004 most of the links from new countries target users in Malaysia, Singapore and Thailand. More or less the same happens in 2005. In 2006 this dynamic changes and we can see that many links appear in new countries, and a huge proportion of these links target users into the



(a) 2003 (b) 2004

(c) 2005 (d) 2006

(e) 2007 (f) 2008

(g) 2009 (h) 2010

Figure 1.3: Distribution of myGamma users by year in which they joined the network.

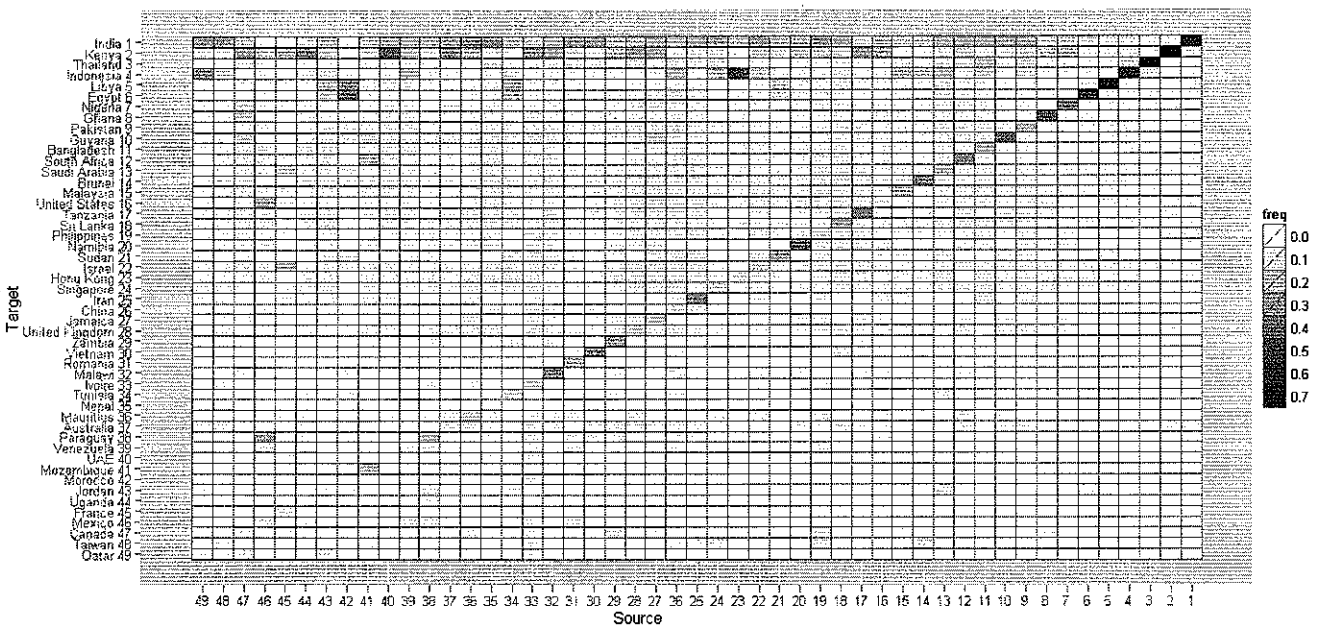


Figure 1.4: Each column shows the distribution of the target's country given the source's country for friendship links. Users that accessed during August 2010. Countries with more than 20 active users in August 2010.

same country. Since 2006 to 2009 we can see that the country–community structure is stronger but a huge number of links target users in India, Indonesia, Kenya, Thailand and the US.

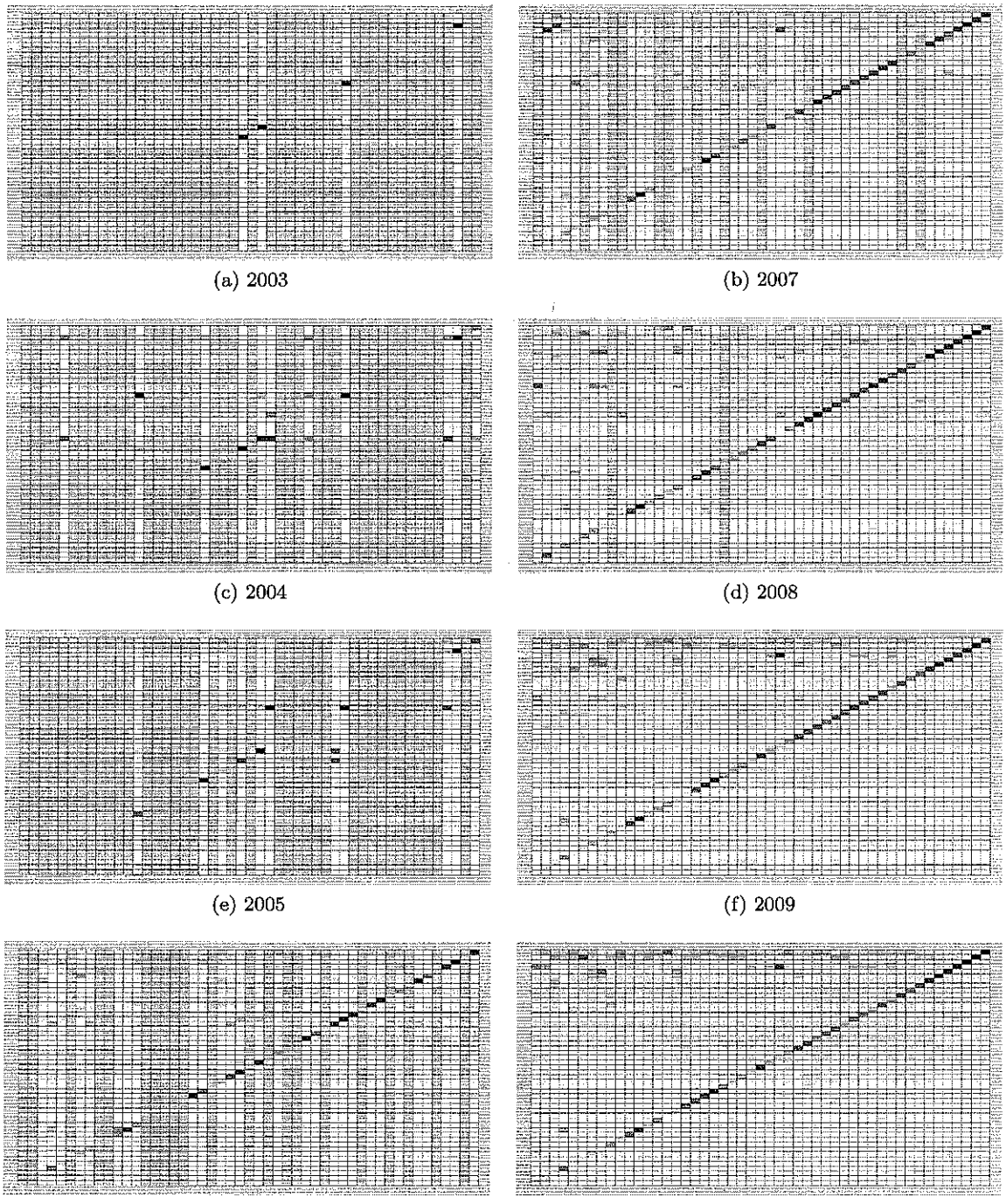
Among all these years there are some interesting patterns that we mention below, even though for the reader it is difficult to check. The graphics in their original sizes can be obtained from the author upon request. The patterns that we mention below correspond to dark squares out of the diagonal of the fluctuation plots in Figure 1.5. The first interesting fact is that in 2007 and 2009 most of the links from Hong Kong targeted users in Indonesia. Also for this year most of the links from Canada and Taiwan targeted users in Thailand and Indonesia. In 2008 most of the links from Qatar targeted users in Saudi Arabia. In 2010 most of the links of new users in Uganda targeted users in Kenya. Finally, also in 2010 most of the links from Nepal targeted users in India.

We can see that it is a reasonable hypothesis to say that in some countries most of the users of myGamma are immigrants.

1.6 Mean of User–Distribution of Friendships

The plots presented in Figure 1.4 and Figure 1.5 may have some problems for countries with small number of users. Whenever the number of users is small and there exists one user with a high number of links targeting users in the same country, the plot is going to be dominated by this information. In order to explore further the hypothesis presented in previous sections we would need to control the effect of users with large numbers of friendships. In order to do this, for countries with more than 20 and less than 100 active users during August 2010, we take for each country the mean of the user–distributions of friendships, i.e. for each user we compute the distribution of friendships and later we take the average of these distributions for all the users within each country. These mean distributions are presented in the columns of Figure 1.6.

If we compare Figure 1.4 and Figure 1.6 we can see that most of the links from UAE (40) target users in



(a) 2003 (b) 2007

(c) 2004 (d) 2008

(e) 2005 (f) 2009

(g) 2006 (h) 2010

Figure 1.5: For each year, each column shows the distribution of the target's country given the source's country for friendship links. Each plot shows the information of the users that joined myGamma during the corresponding year. Countries with more than 20 active users during August 2010.

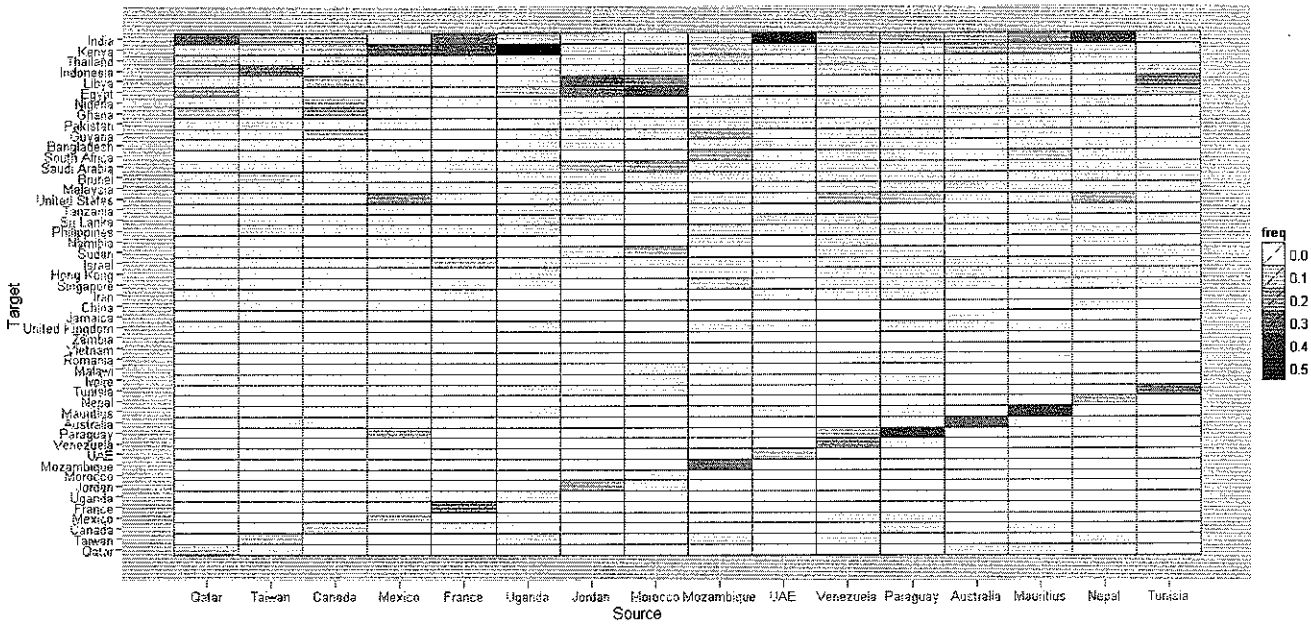


Figure 1.6: Each column shows the mean user–distribution of the target's country given the source's country for friendship links. Users that accessed during Aug 10. Countries with more than 20 and less than 100 active users.

Kenya, however, the mean user–distribution targets mostly users in India. From Figure 1.6 we also can see that for many countries the mean proportion targeting users in the same country increases, but for some countries the highest mean proportion of targets is still located in another country. A clear case is Uganda where the mean user–distribution targets Kenya the most.

Chapter 2

Modeling Latent Relationship Strength Using Count Interaction Data

2.1 Introduction

We propose a model to measure "latent relationship strength" between users of online social networks. The motivation to do so is to have a more reliable measure of the relationship between users besides declared links. The proposed model uses different types of interaction between users such as messages and chats. The intuition for the model is that the more interaction the larger the latent relationship strength. Xiang et al. (2010) presents a model that has a similar hierarchical structure to the one proposed here, but unlike their proposal, we use an approach inspired in the exploration of the data.

In order to measure latent relationship strength (LRS) we propose a model that aims to fit marginally the distribution of the observed interaction between users. We firstly check how well our model works for a small dataset. In the previous chapter we saw that an important number of declared friendships appear within each singular country, so taking the users in certain country might be a good subset to study. We also saw that the myGamma network has been loosing users in South Africa. That is the reason understanding the dynamic of the network in that country is important. We focus on the users in that country from now on. Although our model is static so far, future dynamic versions might help to the understanding of the evolution of the network. We also choose users that were using the networking service since before 2010 and who were active during November 2010 in order to have a stable group of users. We take the interaction during November 2010 represented in messages and chats among pairs of users with declared friendships.

2.2 General Intuition of the Model

In Figure 2.1 we present the histograms of messages sent by chat (just *chats* from now on) and regular messages (just *messages* from now on) among the South Africa group of users, along with the corresponding scatter plot. We can see that there are only a few pairs of users with large amounts of interaction, whereas most of the interaction is null or pretty small. Our idea to model the LRS takes into account this structure of the data. We consider that only a few declared friendships have large LRS and most of the declared friendships have small or null LRS.

The general idea of the model states that each of the K interaction variables from user i to j, X_{ij}^k, such as chats and messages, have some counting distribution F_k conditioning on the value of the LRS. If we call λ_{ij} the LRS from users i to j, we model λ_{ij} using some distribution F_λ defined in the non–negative reals, with a monotonically decreasing pdf such as in the gamma family with a shape parameter lower or equal to one. Thus,

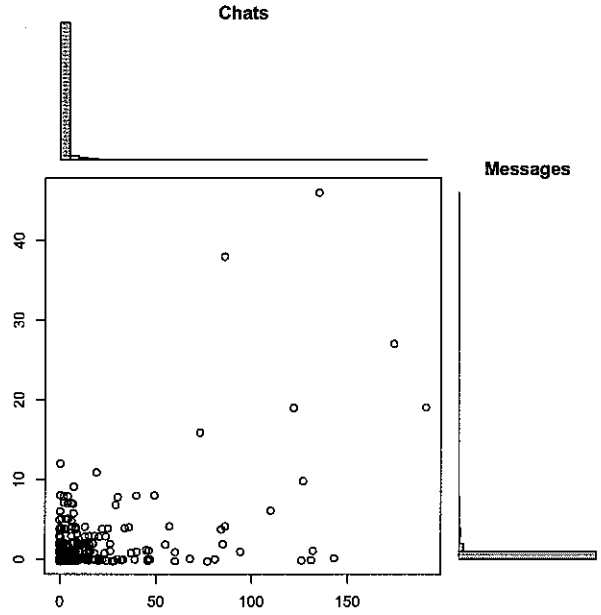


Figure 2.1: Interaction during November 2010 between the 1734 pairs of users with declared friendships in South Africa.

the general idea of the model can be represented in a hierarchical structure as

$$\begin{aligned} \lambda_{ij} &\sim F_\lambda \\ X_{ij}^k|\lambda_{ij} &\sim F_k \end{aligned} \tag{2.1}$$

Further exploration of the data however suggests that the distribution of the LRS might depend on covariates. For instance in Figure 2.2 we can see that the large amounts of interaction occur from males to females and viceversa. The amounts of interaction within gender are very small. According to this observation, the distribution of the LRS between genders might have a heavier tail than within genders. Consequently, we should allow our model to take into account covariates in order to measure the latent relationship strength.

Finally, even taking into account all the relevant covariates to model the LRS, there might be some remaining variability in the interaction variables that is not possible to explain only using the LRS. This might be the case for users with large propensity to interact via certain applications of the online social network. Naturally the propensity to interact using certain features of the network is differential. Thus, we should also take into account covariates that may explain the propensity of users to interact, besides the LRS, such as the number of people each user interacts with.

2.3 Parametrization of the Model

In this section we provide a parametrization of the model presented in equation (2.1). Although we could use other modeling options, our approach seems appealing due to its simplicity.

Let us use the latent indicator variable δ_{ij} as

$$\delta_{ij} = \begin{cases} 1 & \text{if the relationship strength between users } i \text{ and } j \text{ is non zero,} \\ 0 & \text{otherwise} \end{cases}$$

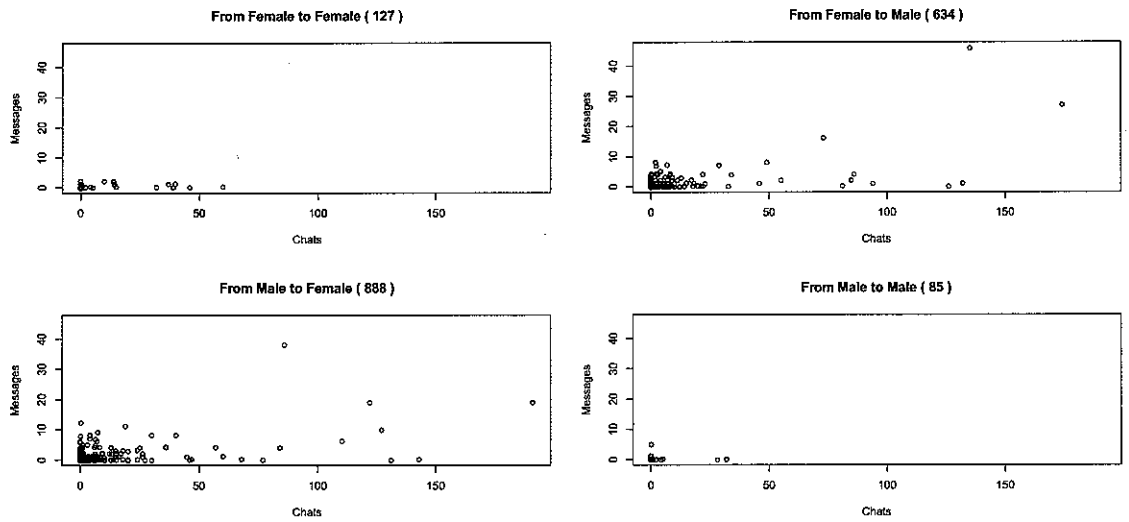


Figure 2.2: Messages vs. chats by the cross–classification of sender's and receiver's gender.

where

$$\delta_{ij} \sim Bernoulli(\pi)$$

We introduce this variable in order to enable the model to handle extra zero inflation of the interaction variables. Let λ_{ij} represent the latent relationship strength from users i to j. We model λ_{ij} as

$$\lambda_{ij}|\delta_{ij}, y_{ij} \sim \begin{cases} Gamma(\alpha, \alpha/\mu_{ij}) & \text{if } \delta_{ij} = 1, \\ 0 & \text{if } \delta_{ij} = 0 \end{cases} \quad \log \mu_{ij} = y_{ij}^T \beta \tag{2.2}$$

where y_{ij} is a vector of covariates comparing information between users i and j, such as gender, location, race or age. Also in (2.2), $\mu_{ij} = E(\lambda_{ij}|\delta_{ij} = 1, y_{ij})$.

Let X_{ij}^k, $k = 1, \ldots, K$ represent the observed interaction variables (chats, messages, comments, etc.) from users i to j. Given the latent relationship strength we model the interaction variables as

$$X_{ij}^k|\lambda_{ij}, z_{ij}^k \sim \begin{cases} Poisson(\gamma_{ij}^k \lambda_{ij}) & \text{if } \lambda_{ij} > 0, \\ 0 & \text{if } \lambda_{ij} = 0 \end{cases} \quad k = 0, \ldots, K; \;\; \log \gamma_{ij}^k = {z_{ij}^k}^T \eta^k \tag{2.3}$$

and z_{ij}^k are covariates that explain the intrinsic propensity for each pair of users to interact via the variable k. For instance, Xiang et al. (2010) propose to use the specific kth interaction outdegree for user i, i.e. the number of people user i interacts with via the kth interaction variable. In this document we use data from previous months to compute this measure in order to avoid using a function of the response data as a covariate (see Section 2.6 for a discussion).

Note from (2.3) that according to the model $E(X_{ij}^k|\lambda_{ij}, z_{ij}^k) = \gamma_{ij}^k \lambda_{ij} = \exp({z_{ij}^k}^T \eta^k)\lambda_{ij}$, or

$$E(X_{ij}^k|\lambda_{ij}, z_{ij}^k)/\lambda_{ij} = \exp({z_{ij}^k}^T \eta^k), \text{ if } \lambda_{ij} > 0$$

which states clearly that we model the variability of the interaction variables after controlling by the LRS, or in other words, we model the expected interaction rate by units of LRS.

2.3.1 Identifiability Constraint

As stated, the above model is unidentifiable if we allow arbitrary intercepts in the linear predictors of both levels of the model. Let us take $\mu_{ij} = \exp(\beta_0)\exp(\sum_{r>0} y_{ij}^r \beta_r) = \mu^0 \mu_{ij}^{\{-0\}}$, where $\mu^0 = \exp(\beta_0)$. Also, taking $\gamma_{ij}^k = \exp(\eta_0^k)\exp(\sum_{s>0} z_{ij}^{sk}\eta_s^k) = \gamma^{k0}\gamma_{ij}^{k\{-0\}}$, where $\gamma^{k0} = \exp(\eta_0^k)$. We can see that the mean of $X_{ij}^k|\lambda_{ij}, z_{ij}^k$ has a conditional distribution

$$\gamma_{ij}^k \lambda_{ij}|\delta_{ij} = 1, y_{ij}, z_{ij}^k \sim Gamma(\alpha, \alpha/\gamma_{ij}^k \mu_{ij}), \text{ for } k = 0, \ldots, K \tag{2.4}$$

where $\gamma_{ij}^k \mu_{ij} = \gamma^{k0}\mu^0 \gamma_{ij}^{k\{-0\}} \mu_{ij}^{\{-0\}}$. Hence, we have μ^0 and $\gamma^{k0}, k = 0, \ldots, K$, as free parameters from which taking infinite combinations of values we could obtain the same $K+1$ distributions in (2.4). Thus, we fix $\mu^0 = 1$ or in other words, we fit a model without intercept for $\log \mu_{ij}$ in (2.2). Note that this approach also implies that the mean LRS is 1 whenever the covariates y_{ij}^r are zero. This might be a good property of the model for comparison between users with different characteristics.

2.3.2 Complete Likelihood

We denote $X_{ij} = (X_{ij}^1, \ldots, X_{ij}^K)$ as the random interaction vector and $x_{ij} = (x_{ij}^1, \ldots, x_{ij}^K)$ as a particular observation. Under the assumption that the observed interaction variables are conditional independent given the friendship levels, we obtain the joint distribution of the complete data for users i and j as

$$p(X_{ij}, \lambda_{ij}, \delta_{ij}) = \left[\frac{\pi^{\delta_{ij}}}{\Gamma(\alpha)\lambda_{ij}}\left(\frac{\alpha\lambda_{ij}}{\mu_{ij}}e^{-\lambda_{ij}/\mu_{ij}}\right)^{\alpha}\prod_{k=0}^{K}\left(\frac{e^{-\gamma_{ij}^k\lambda_{ij}}(\gamma_{ij}^k\lambda_{ij})^{X_{ij}^k}}{X_{ij}^k!}\right)\right]^{\delta_{ij}}\left[(1-\pi)(1-\delta_{ij})\prod_{k=0}^{K} I(X_{ij}^k = 0)\right]^{1-\delta_{ij}} \tag{2.5}$$

from which we easily derive the particular log–likelihood for the interaction between users i and j as

$$\begin{aligned}&\log L_{ij} \propto \\ &\delta_{ij}\left[\log\pi - \log\Gamma(\alpha) + \alpha\left(\log\alpha + \log\lambda_{ij} - y_{ij}^T\beta - \lambda_{ij}\exp(-y_{ij}^T\beta)\right) - \lambda_{ij}\sum_k \exp(z_{ij}^{kT}\eta^k) + \sum_k x_{ij}^k z_{ij}^{kT}\eta^k\right] \\ &+(1-\delta_{ij})\log(1-\pi)\end{aligned} \tag{2.6}$$

which is obtained noting that δ_{ij} is equivalent to $I(\lambda_{ij} > 0)$.

Under the assumption that the complete data for each pair of users are independent, we can find the complete log–likelihood as $\log L = \sum_{ij} \log L_{ij}$.

2.4 Maximum Likelihood Estimation via an EM Algorithm

In order to estimate the vector of parameters $\Phi^T = (\pi, \alpha, \beta^T, \eta^{0T}, \ldots, \eta^{KT})$ via maximum likelihood estimation we use an EM algorithm.

2.4.1 Expectation Step

Using a vector of estimates $\Phi^{(t)}$ from iteration t, the EM algorithm requires for the E step the computation of

$$E_{\delta_{ij},\lambda_{ij}|X_{ij},\Phi^{(t)}}\left[\log L(\Phi; \delta_{ij}, \lambda_{ij}, X_{ij})\right]$$

However, from (2.6) we can see that the log–likelihood is proportional to a quantity which is a linear function of δ_{ij}, λ_{ij}, $\delta_{ij}\log\lambda_{ij}$ and $\delta_{ij}\lambda_{ij}$, thus we only need to find the expectations of these variables given X_{ij}. In order

to facilitate the notation we drop the index (t) from now on, although it should be clear that the parameters in the expectations are $\Phi^{(t)}$. It is clear that

$$E(\delta_{ij}|X_{ij}=x_{ij}) = P(\delta_{ij}=1|X_{ij}=x_{ij})$$

In order to compute this quantity let us find

$$\begin{aligned} p(\delta_{ij}=1, X_{ij}=x_{ij}) &= \int p(X_{ij},\lambda_{ij},\delta_{ij}=1)d\lambda_{ij} \\ &= \frac{\pi\prod_k(\gamma_{ij}^k)^{x_{ij}^k}}{\Gamma(\alpha)\prod_k x_{ij}^k!}\left(\frac{\alpha}{\mu_{ij}}\right)^\alpha \int e^{-\lambda_{ij}(\alpha/\mu_{ij}+\sum_k\gamma_{ij}^k)}\lambda_{ij}^{\alpha-1+\sum_k x_{ij}^k}d\lambda_{ij} \\ &= \frac{\pi\prod_k(\gamma_{ij}^k)^{x_{ij}^k}}{\Gamma(\alpha)\prod_k x_{ij}^k!}\left(\frac{\alpha}{\mu_{ij}}\right)^\alpha \frac{\Gamma(\alpha+\sum_k x_{ij}^k)}{(\alpha/\mu_{ij}+\sum_k\gamma_{ij}^k)^{\alpha+\sum_k x_{ij}^k}} \end{aligned}$$

and similarly

$$p(\delta_{ij}=0, X_{ij}=x_{ij}) = (1-\pi)\prod_k I(x_{ij}^k=0)$$

from which we obtain

$$P(\delta_{ij}=1|X_{ij}=x_{ij}) = \begin{cases} \pi/\left[\pi+(1-\pi)(1+\mu_{ij}\sum_k\gamma_{ij}^k/\alpha)^\alpha\right] & \text{if } x_{ij}^0=\cdots=x_{ij}^K=0, \\ 1 & \text{otherwise} \end{cases}$$

In order to compute $E(\lambda_{ij}|X_{ij}=x_{ij})$ we proceed as follows

$$\begin{aligned} E(\lambda_{ij}|X_{ij}=x_{ij}) &= E(\lambda_{ij}|X_{ij}=x_{ij},\delta_{ij}=1)P(\delta_{ij}=1|X_{ij}=x_{ij}) \\ &= \frac{\alpha+\sum_k x_{ij}^k}{\alpha/\mu_{ij}+\sum_k\gamma_{ij}^k}P(\delta_{ij}=1|X_{ij}=x_{ij}) \end{aligned}$$

where the first equality is due to the fact that $E(\lambda_{ij}|X_{ij}=x_{ij},\delta_{ij}=0)=0$. From (2.5) we can easily check that $\lambda_{ij}|X_{ij}=x_{ij},\delta_{ij}=1 \sim Gamma(\alpha+\sum_k x_{ij}^k,\ \alpha/\mu_{ij}+\sum_k\gamma_{ij}^k)$ and hence we obtain the second equality above, where

$$\mu_{ij}=\exp\left(y_{ij}^T\beta\right) \text{ and } \gamma_{ij}^k=\exp\left(z_{ij}^{k\,T}\eta^k\right)$$

Finally

$$E(\delta_{ij}\lambda_{ij}|X_{ij}=x_{ij}) = E(\lambda_{ij}|X_{ij}=x_{ij},\delta_{ij}=1)P(\delta_{ij}=1|X_{ij}=x_{ij}) = E(\lambda_{ij}|X_{ij}=x_{ij})$$

and

$$\begin{aligned} E(\delta_{ij}\log\lambda_{ij}|X_{ij}=x_{ij}) &= E(\log\lambda_{ij}|X_{ij}=x_{ij},\delta_{ij}=1)P(\delta_{ij}=1|X_{ij}=x_{ij}) \\ &= \left[\psi\Big(\alpha+\sum_k x_{ij}^k\Big)-\log\Big(\alpha/\mu_{ij}+\sum_k\gamma_{ij}^k\Big)\right]P(\delta_{ij}=1|X_{ij}=x_{ij}) \end{aligned}$$

where $\psi(\cdot)$ represents the digamma function.

2.4.2 Maximization Step

For the M step we need to find

$$\arg\max_{\Phi}\left\{ E_{\delta_{ij},\lambda_{ij}|X_{ij},\Phi^{(t)}}\left[\log L(\Phi;\delta_{ij},\lambda_{ij},X_{ij})\right]\right\}$$

In this case we just need to replace in $\log L$, δ_{ij} by $\delta_{ij}^{(t+1)} = E_{\Phi^{(t)}}[\delta_{ij}|X_{ij}=x_{ij}]$, λ_{ij} and $\delta_{ij}\lambda_{ij}$ by $\lambda_{ij}^{(t+1)} = E_{\Phi^{(t)}}[\lambda_{ij}|X_{ij}=x_{ij}]$ and finally $\delta_{ij}\log\lambda_{ij}$ by $\nu_{ij}^{(t+1)} = E_{\Phi^{(t)}}[\delta_{ij}\log\lambda_{ij}|X_{ij}=x_{ij}]$. We proceed to maximize the log–likelihood $\log L$ as if $\delta_{ij}^{(t+1)}$, $\lambda_{ij}^{(t+1)}$ and $\nu_{ij}^{(t+1)}$ were observed data.

From (2.6) we can see that the values of π, β and $\eta^0,\ldots,\eta^K$ that maximize $\log L$ can be obtained maximizing independently the terms of the log–likelihood that involve each respective parameter. The value of π that maximizes $\log L$ can be found simply by taking derivatives from which we find

$$\pi^{(t+1)} = \sum_{ij}\delta_{ij}^{(t+1)}/n$$

where n denotes the number of pairs of users. In order to maximize $\log L$ with respect to β and $\eta^k, k=0,\ldots,K$, we can just use a Nelder–Mead algorithm to maximize the functions

$$f(\beta) = -\sum_{ij}\left(\delta_{ij}^{(t+1)}y_{ij}^T\beta + \lambda_{ij}^{(t+1)}\exp(-y_{ij}^T\beta)\right)$$

$$g_k(\eta^k) = \sum_{ij}\left(-\lambda_{ij}^{(t+1)}\sum_k \exp({z_{ij}^k}^T\eta^k) + \delta_{ij}^{(t+1)}\sum_k x_{ij}^k {z_{ij}^k}^T\eta^k\right)$$

Thus, $\beta^{(t+1)} = \arg\max_{\beta} f(\beta)$ and $\eta_k^{(t+1)} = \arg\max_{\eta_k} g_k(\eta_k)$. Now, using $\beta^{(t+1)}$ we can find $\alpha^{(t+1)} = \arg\max_{\alpha} h(\alpha)$, where

$$h(\alpha) = -\log\Gamma(\alpha)\sum_{ij}\delta_{ij}^{(t+1)} + \alpha\left(\log\alpha\sum_{ij}\delta_{ij}^{(t+1)} + \sum_{ij}\nu_{ij}^{(t+1)} + f\left(\hat{\beta}^{(t+1)}\right)\right)$$

2.4.3 Starting Values

In order to start the EM algorithm we propose to take

$$\pi^{(0)} = 1 - \sum_{ij}\prod_k I(x_{ij}^k = 0)/n$$

Since π measures the overall probability of having LRS grater than zero, we take $\pi^{(0)}$ as one minus the proportion of pairs of users without any kind of interaction.

We also notice that $X_{ij}^k|\delta_{ij}=1 \sim NegativeBinomial(\alpha,\alpha/(\alpha+\mu_{ij}\gamma_{ij}^k))$ as follows

$$\begin{aligned}\int p(X_{ij}^k|\lambda_{ij})p(\lambda_{ij}|\delta_{ij}=1)d\lambda_{ij} &= \frac{(\gamma_{ij}^k)^{x_{ij}^k}}{\Gamma(\alpha)x_{ij}^k!}\left(\frac{\alpha}{\mu_{ij}}\right)^{\alpha}\int e^{-\lambda_{ij}(\alpha/\mu_{ij}+\gamma_{ij}^k)}\lambda_{ij}^{\alpha-1+x_{ij}^k}d\lambda_{ij}\\ &= \binom{x_{ij}^k+\alpha-1}{x_{ij}^k}\left(\frac{\alpha}{\mu_{ij}\gamma_{ij}^k+\alpha}\right)^{\alpha}\left(\frac{\mu_{ij}\gamma_{ij}^k}{\mu_{ij}\gamma_{ij}^k+\alpha}\right)^{X_{ij}^k}\end{aligned}$$

Since $X_{ij}^k|\delta_{ij}=0$ is distributed as the zero constant, and $E(X_{ij}^k|\delta_{ij}=1)=\mu_{ij}\gamma_{ij}^k$, it is easy to see that we are modeling marginally

$$E(X_{ij}^k)=\pi\mu_{ij}\gamma_{ij}^k=\exp\left(\log\pi+y_{ij}^T\beta+{z_{ij}^k}^T\eta^k\right)$$

Thus, in order to obtain initial values $\alpha^{(0)}$, $\beta^{(0)}$ and $\eta^{k(0)}$, $k=1,\ldots,K$, we can fit a negative binomial generalized linear model (see Hilbe, 2011) for each X_{ij}^k using the log link and $\log\pi^{(0)}$ as an offset, as well as y_{ij} and z_{ij}^k as covariates. From this procedure we obtain starting values $\eta^{k(0)}$ for each of the K different negative binomial models. However, note that from this procedure we would obtain different values of $\beta^{(0)}$ and $\alpha^{(0)}$, for each different k. We call those different values $\beta_k^{(0)}$ and $\alpha_k^{(0)}$. We take $\beta^{(0)}=\sum_k\beta_k^{(0)}/K$ and $\alpha^{(0)}=\sum_k\alpha_k^{(0)}/K$.

2.5 Application

In this section we present two implementations of our method using different options of comparison data for the set of myGamma users described in Sections 2.1 and 2.2. For the first implementation we take the interaction variables X^0 as number of chat messages (simply *chats* from now on) and X^1 as number of regular messages (simply *messages* from now on). In the second approach we take X^0 as number of chat sessions and X^1 as number of regular messages. Note that the number of chat sessions between a pair of users is a symmetric measure of interaction, which implies a loss of information since it ignores how involved were the users in the conversations. The motivation for including chat sessions as one of the interaction variables comes from the fact that its distribution is less dispersed, which makes its modeling easier.

2.5.1 Chats and Messages

We take the interaction variables X^0 as chats and X^1 as messages during November 2010. We take gender comparisons as covariates to model the mean latent relationship strength: female–female, female–male, male–female (male–male is the baseline for comparison). Also, as covariates to control the user's propensity to interact via chats and messages we take the chats and messages outdegrees from the interaction variables during the period from July to October 2010.

After fitting the model we obtain $\hat{\pi}\approx 1$ which indicates that for these data we did not require the extra level of the model to handle extra zero inflation. The zero inflation is captured by the shape parameter of the gamma distribution of the LRS, which is estimated to be $\hat{\alpha}=0.07$. This small value makes the gamma distribution to have a large mass near zero. The estimated model for the mean LRS is

$$\hat{\mu}_{ij}=\exp\left(0.64I[female(i)=female(j)]+0.86I[female(i)=male(j)]+0.69I[male(i)=female(j)]\right)$$

from which we obtain the mean LRS for the different gender comparisons as $\hat{\mu}_{m-m}=1$ (baseline), $\hat{\mu}_{f-f}=1.89$, $\hat{\mu}_{f-m}=2.37$ and $\hat{\mu}_{m-f}=2.00$. The interpretation of these values is pretty straightforward. For instance, we estimate that the mean LRS from females to males is 2.37 times the mean LRS from males to males. The remaining mean LRS can be read in a similar fashion.

In Figure 2.3 we present a histogram of the estimated LRS, where we can see that its shape describes our initial intuition about the distribution of LRS.

For the models that control the user's propensity to interact via chats and messages we obtain

$$\hat{\gamma}_{ij}^0=\hat{E}(chats_{ij}|\lambda_{ij},chatoutdegree_i)/\lambda_{ij}=\exp\left(-0.40+0.06chatoutdegree_i\right)$$

and

$$\hat{\gamma}_{ij}^1=\hat{E}(messages_{ij}|\lambda_{ij},msgsoutdegree_i)/\lambda_{ij}=\exp\left(-2.10+0.09msgsoutdegree_i\right)$$

Thus, according to these two models, the expected rate of chats gets multiplied by $\exp(0.06)=1.05$ when the chats outdegree is one unity larger. Similarly, the expected rate of messages gets multiplied by $\exp(0.09)=1.09$

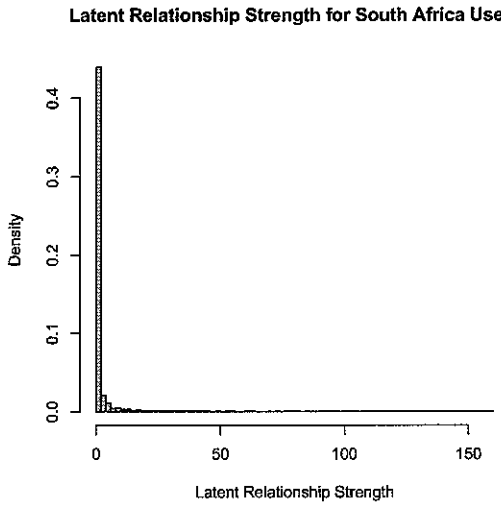


Figure 2.3: Histogram of the estimated latent relationship strength for users with declared friendships.

when the messages outdegree is one unity larger. Also, according to the model, the expected number of chats is $\exp(-0.40 + 2.1) = 5.47$ times the expected number of messages, after controlling by the LRS and the chats and messages outdegrees.

Finally, we explore the marginal fit of the model to the distribution of the interaction variables. According to the model

$$X_{ij}^k | y_{ij}, z_{ij}^k \sim \pi NegativeBinomial(\alpha, \alpha/[\alpha + \mu_{ij}\gamma_{ij}^k])$$

Thus we compute the expected counts $\hat{P}(X_{ij}^k = x) = \sum_{ij} \hat{P}(X_{ij}^k = x | y_{ij}, z_{ij}^k)$ and we compare these expected counts with the observed distribution of each interaction variable. In Figure 2.4 we present both the estimated marginal frequencies and the observed interaction frequencies. We can see that the marginal fit of the model to the observed interaction distributions is not very good, specially for chats. In the next section we explore the fit of our model using chat sessions.

2.5.2 Chat Sessions and Messages

In this section we take the interaction variables X^0 as chat sessions and X^1 as messages during November 2010. Consequently we take the outdegrees for chat sessions and for messages during July to October 2010 as the covariates to control the user's propensity to interact via chats sessions and messages, respectively. For these interaction data we also obtain $\hat{\pi} \approx 1$ and $\hat{\alpha} = 0.16$. The estimated model for the mean LRS is

$$\hat{\mu}_{ij} = \exp\left(0.99 I[female(i) = female(j)] + 1.58 I[female(i) = male(j)] + 1.48 I[male(i) = female(j)]\right)$$

The estimated models that control the user's propensity to interact via chats sessions and messages are

$$\hat{\gamma}_{ij}^0 = \hat{E}(chatsessions_{ij} | \lambda_{ij}, chatsesoutdegree_i)/\lambda_{ij} = \exp\left(-3.10 + 0.05 chatoutdegree_i\right)$$

and

$$\hat{\gamma}_{ij}^1 = \hat{E}(messages_{ij} | \lambda_{ij}, msgsoutdegree_i)/\lambda_{ij} = \exp\left(-2.62 + 0.05 msgsoutdegree_i\right)$$

These models can be interpreted as in Section 2.5.1. In Figure 2.5 we present the fit of our model to the marginal distribution of chat sessions and messages. This figure shows that the marginal fit of our model to these interaction data is reasonable.

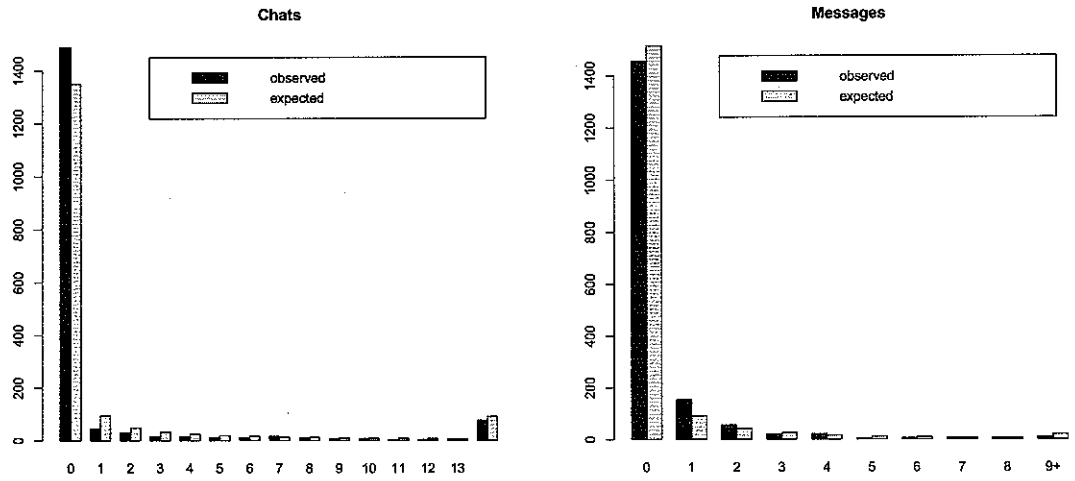


Figure 2.4: Checking the marginal fit of the model for chats and messages.

2.6 Discussion and Future Work

We believe our approach to measure LRS provides improvements over the proposal of Xiang et al. (2010). Their approach presents the following characteristics:

- Each interaction variable is dichotomized into two categories that measure whether the interaction is null or not. This approach treats all the non–null interaction in the same way and, as a consequence, the range of values of LRS that the method can recover becomes limited.
- The covariates to measure the user's propensity to interact are the outdegrees of the interaction variables for the same period of time for which the LRS is being computed. Thus, since the interaction's outdegree is a function of the interaction variable, a function of the response variable is being used as an explanatory covariate.
- F_λ is parameterized using a normal distribution. Although it is a computationally appealing approach, the election of a normal distribution to model LRS does not reflect the characteristics of the phenomenon under study.

We presented a model that aims to measure latent relationship strength in online social networks directly using count interaction data. Our approach aims to check the marginal fit of the model to the observed interaction data. Furthermore, covariates can be incorporated into the model in order to explain the variability of the LRS by different characteristics of pairs of users. Our approach also models the LRS according to the intuition provided by the data. Nevertheless, the proposed model seems to require more calibration in order to capture reasonably well the variability of the data. Further efforts have to be made in order to clarify how to assess its fit.

Both Xiang et al. (2010) and our approach are criticizable in a number of ways that we will address in a future extension of this work:

- λ_{ij} and λ_{ji} are modeled independently. This assumption is too simplistic since we would expect to see a large correlation among the two LRS arising from a pair of users.
- There is lack of assessment of the fit of the models.

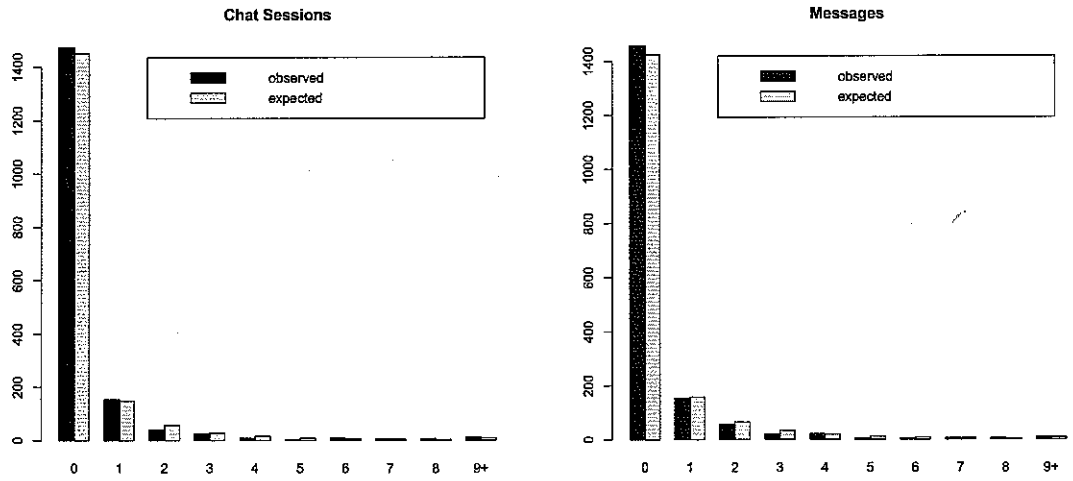


Figure 2.5: Checking the marginal fit of the model for chat sessions and messages.

- Covariates/features selection is not addressed.

In a future version of this work we will allow the model to take into account the correlation of LRS for pairs of users following some of the ideas presented in Holland and Leinhardt (1981). In order to assess the fit of the model we will implement the ideas presented in Hunter et al. (2008) and we will explore ways to do model selection in order to determine which covariates should be included in the final model.

Acknowledgements

Cosma Shalizi, Mark Schervish, and Andrew Thomas made helpful suggestions that contributed to improvements in this work. This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

Bibliography

Hilbe, J. M. (2011). *Negative Binomial Regression.* Cambridge University Press, second edition.

Holland, P. W. and Leinhardt, S. (1981). An Exponential Family of Probability Distributions for Directed Graphs. *Journal of the American Statistical Association*, 76(373):33–50.

Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008). Goodness of Fit of Social Network Models. *Journal of the American Statistical Association*, 103(481):248–258.

Leung, C. W., Lim, E.-P., Lo, D., and Weng, J. (2010). Mining Interesting Link Formation Rules in Social Networks. The 19th ACM Conference on Information and Knowledge Management (CIKM2010).

On, B.-W., Lim, E.-P., Jiang, J., Chua, F. C. T., Nguyen, V.-A., and Teow, L.-N. (2010). Messaging Behavior Modeling in Mobile Social Networks. Symposium on Social Intelligence and Networking (SIN-10), in conjunction with Second IEEE International Conference on Social Computing (SocialCom2010).

Xiang, R., Neville, J., and Rogati, M. (2010). Modeling Relationship Strength in Online Social Networks. The 19th International World Wide Web Conference (WWW), 2010.