Statistical Analysis and Anomaly Detection of SMS Social Networks

Research-in-Progress

Bin Zhang
Heinz College and iLab
Carnegie Mellon University
binzhang@cmu.edu

Liye Ma
Robert H. Smith School of Business
University of Maryland
liyema@rhsmith.umd.edu

Ramayya Krishnan
Heinz College and iLab
Carnegie Mellon University
rk2x@cmu.edu

Abstract

Social network analysis has attracted intensive interests by researchers from multiple disciplines. However most of the existing work is descriptive nature, and statistical network analysis remains an active area of research. In this paper, we model and study two facets of the social networks in short message services (SMS). One is the structure of the contact networks of mobile users, the other is users' messaging behavior pattern. We want to account for the heterogeneity in behavior so that to identify abusive usage such as spamming through the study. We use power-law mixture model to capture community formation behaviors, the first facet, and use Poisson-panel mixture models to uncover abnormal behaviors in text messaging. Our results show heterogeneity of the consumers' sending behavior, also there are two major types of community formation behavior in SMS network.

Keywords: social network, anomaly detection, parameter estimation, power-law

Introduction

Short message services (SMS) has become an important method of communication in many countries, mainly because the lower price compared to cellular phone calls. For example, a 450 minutes calling plan at Verizon costs $39.99, while an unlimited text messaging plan is only $20.00 (Verizon 2011). Text messaging also has the advantage that it can be sent and replied to in time, while phone calls may not be answered at any time in some situations, such as during a meeting. All these benefits cause a soar in messaging usage. The number of SMS messages sent increased from 81 billion in 2005 to 2.1 trillion in 2010 (6,933 messages/subscriber/year; by Cellular Telecommunications Industry Association, 2010). As one of the major socially interactive technologies, text messaging is beginning to redefine the formation of social networks, especially among youth (Bryant et al. 2006). By offering real-time and inexpensive communication, SMS messaging facilitates the formation of social networks and helps it evolve over time. These text messaging social networks may affect the offline real social networks to which individuals belong. Consequently, they have great potential to help uncover important social behaviors that cannot be done before.

Because of the pervasiveness of SMS, marketers also have leveraged such technology to make it an important channel for advertising. Because of its great reach, together with its low cost, SMS became the major medium of mobile advertising (De Reyck and Degraeve, 2003). However, mobile marketing has also brought the serious problem of SMS spamming, which is unsolicited commercial messages through the use of SMS. Currently 80% of cellular phone subscribers have received spam. Spamming could result in waste of bandwidth, memory overflow of handsets, service disruption, and consequently increased cost

and significant loss of revenue (Ong 2010), thus identifying the abusive behavior is crucial for maintaining healthy functioning of the service. Although a spammer could be potentially identified from sender behavior – a sender who sends a large number of messages in a short time has a high possibility to be a spammer. However, there are challenges from such a problem. First, the definition of spamming behavior – there is no accurately defined threshold to differentiate spamming. Second, a network may consist of heterogeneous groups of people with heterogeneity in behaviors, such as some people being heavy users while others are light users. If we simply use a traditional outlier detection method, we may risk including many heavy users. Thus, we want to use Business intelligence (BI) techniques such as statistics modeling to identify potential spammers.

BI is becoming a key enabler for increasing business value and enterprise performance (Waston and Wixom 2007). In general, BI is a collection of technologies, such as data mining, statistical modeling, and machine learning etc, helping enterprises perform data querying and predictive analytics. It is current widely implemented in organization IT infrastructures (Anderson-Lehman et al. 2004). In this paper, we want to use BI techniques to perform social networks analysis and identify potential anomalies.

Social network analysis has drawn tremendous interest in fields such as information systems, organizational sciences, marketing, and computer science. Researchers are especially interested in relationships formed among people with common characteristics, behavior or interest. There is a rich literature discussing patterns in social networks, and the knowledge has been applied to anomaly detection. However existing work is largely of descriptive nature while in-depth statistical analysis is still in the early stages. Existing work using network data is more often of descriptive nature, while rigorous statistical network analysis remains an active area of research. In this paper, we study the group formation and messaging behavioral patterns in a network of SMS users. Two characteristics are of interest in this context. The first is the structure of the contact networks of mobile users, and the second is the patterns of users' messaging behavior. When the network is modeled as a dynamic graph, the former is reflected from the (weighted) degree distribution of nodes, while the latter is reflected from patterns of the time stamps and destinations of messages. Both can be modeled with appropriate probabilistic distributions. Literature has characterized degree distributions as obeying “power law” (Christos et al. 1999) while the traffic of each sender and receiver can be considered as given number of events occurring in a fixed interval of time thus following Poisson distributions.

Finite mixture models are an ideal choice for modeling heterogeneous behaviors. By using mixture models, we can uncover different patterns of typical behaviors from a messaging and group formation perspective. Some of them will be normal, e.g. some using it heavily, while other using it lightly. Other patterns, however, could be considered as abnormal, e.g. a person who sent messages to 10 thousand people in an hour must be a robot code. Learning a mixture model can both uncover the detail of different types of normal and abnormal behaviors, and assign individual users to the corresponding behavioral pattern. We want to model the group formation behaviors as a mixture of power-law distributions. Through careful development, we reduce this power-law mixture into a mixture of Gaussians. Meanwhile, the messaging behaviors are modeled as a mixture of Poisson distributions. Our model extends the standard Poisson mixture models to both handle the panel data and be robust in the presence of anomalies. Expectation-maximization algorithm is used for estimation, and we confirm the identification and convergence properties of the two models based on existing literature. As the number of different behavioral patterns is not known in advance, we use the Bayesian information criterion for model selection.

The paper is organized as follows: In the section Literatures, we give background knowledge of social networks. The detailed theoretical development of our models, the corresponding estimation methods, and their properties are presented in Section Methods. Results from our statistical analysis are presented in Section Results. Conclusion and future work are suggested in the last section.

Literatures

Several streams of work are related to our paper. The first stream, from the application perspective, is the social network literature. Social networks have drawn tremendous interest from information system (IS) researchers recently (Oinas-Kukkonen et al. 2010). Actually such interest may trace back to technology diffusion in the 1990s (Brancheau and Wetherbe 1990; Chatterjee and Eliashberg 1990; Premkumar and

Nilakanta 1994), while the appropriation of social network approaches in IS may just be starting (Agarwal et al., 2008). One of the most common representations of social networks is an actor to actor graph where edges represent ties between entities. The entities are referred as actors. Actors are linked to other actors by social ties, which are created based on relationships among actors. In our case, actors are cellular phone users who send and receive text messages, while ties are connections created by text messages. The concentration of network analysis is the ties among actors. If a tie is between two actors, it is called a dyad, and if between three actors, it is called a triad, which is also called as a triangle in computer science literature. On a larger scale, larger number of actors tied together can be modeled as a group (Wasserman and Faust 1994). Several network structures that particular interest researchers are clique, star and ring, etc.

The second stream, from a theoretical perspective, is the statistical literature on modeling the behaviors of interest. McLachlan and Peel (2000) provide a comprehensive review of Finite Mixture modeling, which refers to other studies, such as Teicher (1960), for certain theoretical properties of specific mixture models. Bishop (2006) has shown the advantage of mixture models and Casella and Berger (1990) also have shown detailed discussions of the probability distributions used in this paper. Furthermore power-law has been identified in different network contexts. Leskovic et al. (2007) found the popularity of blog posts drops with a power law ($\gamma = -1.5$), such result matches Barabasi's (2005) theory of heavy tails in human behavior. The size distribution of cascades (the number of involved posts), follows a perfect Zipf's distribution with a slope of –2. The most popular shapes in blog network were the stars, i.e. single post with several in-links, but none of the citing posts are themselves cited (Leskovec et al. 2007).

The third stream is anomaly detection in network data. Advancement in data collection technology make it easy for us to get data. The large volume data and the existence of special patterns underlining the data, make this problem interesting and relevant to business intelligence research. Anomaly detection has been addressed in large scale datasets, especially categorical data (Das et al. 2008), but not as easily on network data. The data size and the complicated interdependency between actors in the network make the problem harder to solve. Some notable researches addressing anomaly detection in network data, for example, fraud in online auction, are Pandit et al. (2007) and Zhang et al (2008).

Since our data size is large, we have to construct a subpopulation. However, random sampling with respect to these SMS will not preserve the social network's structure at the local level. A more legitimate and efficient method to build one (or more) subpopulation(s) in this situation is snowball sampling. This method was introduced by Coleman (1958) and Goodman (1961). This method has been extended by Salganik and Heckathorn (2004). In their method, an individual sample is formed by randomly selecting a user from the network and returning the connected component containing this user, repeating this on the remaining users until some maximum number of users is attained.

Methods

The SMS users and their messaging behavior are naturally represented as a weighted-graph or network, with a node representing an individual user and a weighted directed edge representing the existence and amount of messages sent from one user to another. In this paper, we are interested in two attributes of the network structure, namely the weight-degree correlation and the messaging pattern of individual nodes. In the subsequent discussion, the terms "node" and "user" are used interchangeably, as are the terms "message count" and "weight".

Modeling Messaging Patterns

Various distributions have been used to model discrete events in statistical literature, such as Poisson distribution, Negative-Binomial distribution, and Geometric distribution. Among them, Poisson distribution is among the most popular to model the occurrences of independent random events. The action of sending an individual message is naturally considered an independent random event, thus Poisson distribution is a good fit for modeling it. The probability-mass-function (pmf) of a Poisson distribution is:

$$P(X = k \mid \lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!} \tag{1}$$

where λ is the rate parameter which measures the intensity of the occurrence (messages sent).

Poisson distribution assumes that the occurrences of different events are independent. This could be a concern in our context: it is reasonable to expect that the messages sent by an individual within a short period are correlated, such as when two users engage in back-and-forth discussion on a specific matter on a certain day. However, for a longer unit time interval, such as one week, this dependence concern will be largely insulated. Consequently, the unit time period is chosen to be a week.

A user's messaging pattern can thus be modeled as a panel of independent Poisson draws. Let T be the total number of time periods, and let $C_{i,1}$, $C_{i,2}$, ... , $C_{i,T}$ be the number of messages sent by the user i in each time period. The probability of observing such a sequence is thus:

$$P(C_{i,1},...,C_{i,T} \mid \lambda) = \prod_{t=1}^{T} P(C_{i,t} \mid \lambda) = \prod_{t=1}^{T} \frac{e^{-\lambda} \lambda^{C_{i,t}}}{C_{i,t}!} \quad (2)$$

where λ is the user's rate parameter.

Modeling Weight-Degree Correlation

Literature has established that many features in social networks follow "power-law". Specifically, literature in McGlohon et al. (2008) has shown that the in(out)-degree and the in(out)-weight of individual nodes in a weighted-graph follow a power law distribution. To study user's messaging sending behavior, we focus on the out-degree/out-weight relationship. Generically speaking, two random variables X and Y follow power-law if $Y \propto X^{r}$, where r is the power-law coefficient or "slope". Let D_i be the out-degree (total number of different users to whom this user sent messages) of a node i, and W_i be its out-weight (total number of messages sent to other users), then a power-law implies:

$$W_i \propto D_i^{r} \quad (3)$$

Mixture Models

All users are not the same. In fact, consumer heterogeneity is the focus of study in social science fields such as marketing. As our data set contains more than one million users, it is reasonable to expect that different behavior patterns exist.

When using Poisson distribution to model messaging patterns, the rate parameter λ captures the intensity of using the service. Some people are heavy users of SMS, for whom we expect a large λ. However, others may be casual users for whom we expect a small λ. Similarly, the slope of the degree-weight power-law measures the concentration of communication. Some user may communicate with few people but quite intensively to each recipient. The slope, r, for such a user is expected to be large. Whereas another user may communicate with many people but lightly with each recipient, for whom a small slope is expected. To accurately describe user behavior, therefore, mixture models are needed.

Poisson-Panel Mixture Model

Let the population be a mixture of M Poisson distributions. Let λ_j (j = 1...M) be the parameter for these Poisson distributions, and p_j (j = 1...M) be the proportion of each Poisson component in the population. Let X_i (i = 1...N) be independent draws from this mixture. For each i, let $C_{i,1}$, $C_{i,2}$, ... , $C_{i,T}$ be the observed count over T time periods. The likelihood of observing a data sequence can then be expressed as:

$$f(C_{i,1}, C_{i,2},..., C_{i,T} \mid \lambda_1...\lambda_M, p_1...p_M) = \sum_{j=1}^{M} p_j P(C_{i,1}, C_{i,2},..., C_{i,T} \mid \lambda_j) \quad (4)$$

Substituting in the likelihood of individual Poisson as stated in equation (2), we then have

$$f(C_{i,1}, C_{i,2},..., C_{i,T} \mid \lambda_1...\lambda_M, p_1...p_M) = \sum_{j=1}^{M} p_j \left(\prod_{t=1}^{T} \frac{e^{-\lambda} \lambda_j^{C_{i,t}}}{C_{i,t}!} \right) \quad (5)$$

The parameter to be estimated are λ_j and p_j (j = 1...M).

Negative Binomial distribution (NBD) is also often used to model a population of Poisson with different rates. Specifically, NBD can be considered as a continuous mixture of Poisson where the rate follows a Gamma distribution. In our study, however, we believe an explicit finite mixture of Poisson is better than NBD model, because Gamma distribution with a shape parameter larger than 1 is uni-modal (Encyclopaedia of Mathematics, Edited by Michiel Hazewinkel, 2002, Springer), as such it cannot model a multi-modal clustering of rates. While the eventual result may indicate the individual rates in our dataset indeed follow a uni-modal distribution, we believe this should not be imposed as an assumption ex-ante. Therefore, we believe mixture model gives us more flexibility than NBD does.

Power-Law Mixture Model

The power-law relationship between node degree and node weight can be expressed as equation (3). To fit this model directly, however, requires non-linear regression, for which a mixture model is not easy to derive. However, equation (2) can be transformed into:

$$\log W_i = a + r \log D_i + \varepsilon \tag{6}$$

Further assume that $\varepsilon \sim \text{Normal}(0,\sigma^2)$, then equation (6) becomes a simple linear regression model, which can be estimated as a Gaussian model.

From this, a power-law mixture model can be specified. Let (W_i, D_i) $(i = 1 \ldots N)$ are drawn independently from a mixture of M power-law distributions. Let (a_j, r_j, σ_j) $(j = 1 \ldots M)$ be the parameter for these power-law distributions, and p_j, $(j = 1 \ldots M)$ be the proportion of each power-law component in the population. The likelihood of observing a data point can then be expressed as:

$$f(W_i, D_i \mid \mathbf{a}, \mathbf{r}, \boldsymbol{\sigma}, \mathbf{p}) = \sum_{j=1}^{M} p_j g(W_i, D_i \mid a_j, r_j, \sigma_j) \tag{7}$$

where $g(\bullet)$ is the distribution of power-law.

In equation (7), the individual power-law likelihood is simply:

$$g(W_i, D_i \mid a_j, r_j, \sigma_j) = \phi\left(\log W_i - a_j - r_j \log D_i; 0, \sigma_j\right) \tag{8}$$

where $\phi(\bullet; \mu, \sigma^2)$ is the density of a Gaussian distribution. The power-law mixture model is thus effectively reduced to a Gaussian mixture model.

Estimation Methods

The incomplete data likelihood functions as expressed in equations (5) and (7) are convex in mixing weight, convex in distribution parameters, but not convex when mixing weight and distribution parameters are considered together. Therefore, standard convex optimization methods are not applicable. An Expectation-Maximization (E-M) algorithm is thus needed. For the Poisson-panel mixture model, the parameters to be estimated are λ_j, and p_j $(j = 1 \ldots M)$. To estimate this using E-M algorithm, in the E-Step, the responsibility of each class for each data point is calculated:

$$w_i(j) = P(X_i \in \text{class } j) = \frac{p_j P(C_{i,1}, C_{i,2}, \ldots, C_{i,T} \mid \lambda_j)}{\sum_{l=1}^{M} p_l P(C_{i,1}, C_{i,2}, \ldots, C_{i,T} \mid \lambda_l)} \tag{9}$$

In the M-step, the parameters are maximized:

$$p_j = \frac{1}{N} \sum_{i=1}^{n} w_i(j) \tag{10}$$

$$\lambda_j = \frac{\sum_{i=1}^{n} w_i(j) \overline{C_i}}{\sum_{i=1}^{n} w_i(j)} \tag{11}$$

where $\overline{C_i}$ is the sample mean of C, the maximum likelihood estimation for the Poisson distribution.

For the power-law mixture model, the estimation is the same as the Gaussian mixture model. In the E-step, the responsibility is calculated:

$$w_i(j) = P\big((W_i, D_i) \in \text{class } j\big) = \frac{p_j \phi\big(\log(W_i - a_j - r_j)\log D_i; 0, \sigma_j\big)}{\sum_{l=1}^{M} p_l \phi\big(\log(W_i - a_j - r_j)\log D_i; 0, \sigma_l\big)} \quad (12)$$

In the M-step, the parameters are maximized:

$$p_j = \frac{1}{N}\sum_{i=1}^{n} w_i(j) \quad (13) \qquad (a_j, r_j, \sigma_j) \sim WLS\big(W_i, D_i, w_i(j)\big) \quad (14)$$

where *WLS*(•) is the weighted-least-square parameter maximization.

Model Selection

The number of mixtures, M, is taken as exogenous in finite mixture estimation, otherwise the model is unidentified. However, this number is not known for our dataset, and in fact is one of the parameters of interest. Model-selection techniques are needed. As this is not a regression problem, cross-validation is not easily applicable. As we intend to find parsimonious models and Bayesian Information Criterion (BIC) penalizes large models, we choose it for model selection. In our actual estimation, we imposed an even higher penalty than BIC does, by multiplying the second term in BIC with a constant that is bigger than one. This is done because we do not feel that the standard BIC penalizes bigger model enough in our case, as Poisson is modeled by only one parameter. Although this multiplied-BIC does not have as rigorous a theoretical foundation as the BIC does, we believe it is better suited for our problem, as otherwise the estimation will provide many more groups than is suitable for interpretation from the social science perspective.

Results

The short-messaging-service (SMS) data over six-month period is obtained from a large Asian telecommunication company (source and raw data confidential). It consists of phone number (hashed) from sender and receiver, date and time of message, message length etc. The entire dataset contains approximately 2.5 million unique users, with more than 200 million messages sent. From this dataset, a subpopulation of 5,000 individual users is selected by snowball sampling with a random start, and all their incoming/outgoing messages are extracted. The estimation is performed on this subset of users.

The number of Poisson mixtures tested and the resulting BIC are reported in Table 1. The best fit is obtained for seven components. The estimates of the rates are reported in Table 2. As is shown in the table, groups have very different message arrival rates, ranging from on average a little more than one message per week, to hundreds of messages a week on average. This is consistent with the standard notion of consumer heterogeneity in social science disciplines such as marketing.

The number of power-law mixtures tested and the resulting BIC are reported in Table 1. The best fit is obtained for a mixture of two components. The estimates are reported in Table 2. The two groups have different sizes, while the first accounting for a little less than 40% of the population, while the second accounting for the rest of 60%.

Analysis

As is shown in the Poisson mixture result table, different users have different rates of sending SMS messages. The group with lowest usage on average sends less than two messages per week. The group with the highest usage, in contrast, sends more than 300 messages per week on average. Most people use SMS rather lightly - more than 80% of the users send on average less than five messages per week. The heavy users, those sending more than 100 messages per week, account for only a little over 1% of the user population.

As is shown in the power-law mixture model result table, users can be roughly partitioned into two groups based on the power-law slope. The first group of users have a power-law slope of 1.48, while the second 1.37. The first group can be described as those who communicate with some people much more intensively than with other people, while the second group are those who communicate with other people

with similar intensities (the difference between the two groups is larger than the numbers seem to suggest, as the power-law coefficient is in the exponent of the equation).

A quick note on the implementation: the estimation of the power-law mixture model was conducted smoothly. This is as expected, as Gaussian mixture models are generally well-behaved. The standard safeguard is put in place to avoid the degenerate case when the variance of a certain group tends to zero (thus likelihood tends to be infinity), by reassigning a larger variance if the variance of any group dips below a threshold during estimation.

The estimation of the Poisson mixture model, however, runs into some complications initially. As expected, the messaging behavior of a few users is not considered normal, they would constantly send out thousands of messages for some time, and then suddenly stop, and resume after some time. In reality, these users are spammers who abuse the system by flooding other users with unwanted messages, usually advertisement. From modeling perspective, Poisson distribution is not a good fit for describing their behavior (in fact, being abnormal, it is doubtful that their behavior can be readily modeled using any standard distribution). This is not a problem theoretically, as the mixture model will simply pick out the best possible fit from the Poisson distribution family (the "Poisson oracle"). Practically, however, this proved to be infeasible, as the likelihood is so low that the density function in R library literally returns likelihood with value of 0, which causes numeric underflow. The mixture model estimation procedure is therefore implemented with an explicit workaround: when a node is considered unlikely to be in any mixture component, it is not assigned to any and will be ignored in the estimation. The estimation works out well after this workaround is implemented.

Interestingly, this workaround achieves another objective - anomaly detection. Points excluded in this estimation are considered abnormal. During estimation, two such users are identified as being abnormal. A visual inspection of their specific messaging patterns seems to confirm the suspicion. The message counts over time for these two users are shown in Figure 1, where y-axis represents the logarithm of number of messages sent. The chart strongly suggests that they are spammers. In practice, this information can be valuable for the telecom company, which can proceed to confirm the anomaly of their behavior (e.g. through inspecting the message content) and take counter measure to maintain the health of the system.

Conclusion

In this paper we presented our preliminary study about people's SMS messaging and group formation patterns in the context of a social network. We used a Poisson distribution to model message-sending behavior and power law distribution to model the relationship between the number of connections and the number of messages of individual users. Since the consumer behaviors are heterogeneous, we use mixture models to uncover different patterns that exist in the network. Our result confirmed our hypothesis about the heterogeneity of the consumers' SMS sending behavior. Using BIC for model selection, we find that consumers are best categorized into seven types, ranging from using SMS occasionally to sending several hundred messages per week. Such analysis also helps us identify anomaly, as three users in our sample are found to have abnormal behavior. Also, the class of consumers who send messages in the hundreds per week ($\lambda = 303$) also warrants closer scrutiny. Combining class membership with the analysis of those users' messaging behavior over time, they may or may not be determined as spammers. Separately, the results from power-law distribution tell us there are two major types of group formation behavior, one forming fewer concentrated connections while the other forming wider connections with lower communication intensity. However, the difference in these two types is not as dramatic as that among different message sending behavioral types.

To make our study more comprehensive and robust, we recommend future study using the whole dataset, as there are millions of users in the whole dataset, which may have more behavioral patterns than uncovered in our sample. Also, we may test alternative models of message-sending behavior and pick the one with the best performance. Furthermore, we wish to investigate how local group changes across time to further help us detect anomalies.

Table 1. Number of Poisson Mixtures and BIC scores

Poisson				Power-law			
m	*BIC Score*	*m*	*BIC Score*	*m*	*BIC Score*	*m*	*BIC Score*
2	−709049	6	−287216	1	−4366.99	4	−4370.25
3	−397699	7	−284592*	2	−4326.31*	5	−4408.50
4	−313219	8	−290543	3	−4380.27	6	−4346.68
5	−284050						

m = number of mixtures

Table 2. Mixture Model Estimation

Poisson						Power-law				
c	*λ*	*Proportion*	*c*	*λ*	*Proportion*	*c*	*a*	*r*	*σ*	*Proportion*
1	1.2	64.7%	5	61.5	2.5%	1	0.078	1.48	0.70	37.3%
2	4.2	18.9%	6	139.6	1.0%	2	0.398	1.37	1.32	62.7%
3	12.0	7.9%	7	303.0	0.3%					
4	27.2	4.5%								

c = number of components

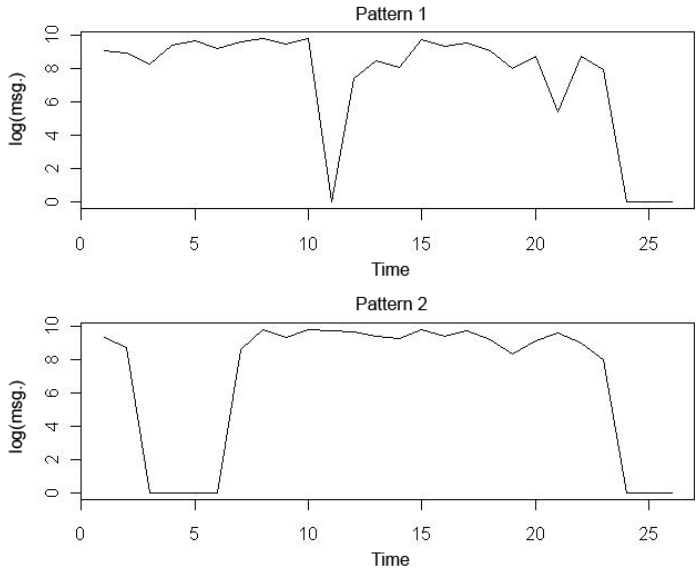

Figure 1. Abnormal messaging patterns

Acknowledgements

This work was supported in part by AT&T and the iLab at Heinz College, Carnegie Mellon University. Dou-Yan Yang gave suggestions to improve this manuscript.

References

Anderson-Lehman, R., Watson H. J., Wixom, B. H., Hoffer, J. A. (2004). Continental Airlines Flies High with Real-Time Business Intelligence. *MIS Quarterly Executive*, Dec., pp. 163–176.

Barabasi, A. -L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, (435:207).

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer.

Brancheau, C. J. and Wetherbe, C. J. (1990). The adoption of spreadsheet software: Testing innovation diffusion theory in the context of end-user computing. *Information Systems Research*, (1:2), pp. 115–143.

Bryant, A. J., Sanders-Jackson, A., and Smallwood, A. M. K. (2006). IMing, text messaging, and adolescent social networks. *Journal of Computer-Mediated Communication*, (11:2).

Casella, G. and Berger, R. L. (1990). *Statistical Inference*. The Wadsworth and Brooks/Cole Statistics/Probability series. Brooks/Cole Publishing. Co, Pacific Grove, CA.

Chatterjee, R. and Eliashberg, J. (1990). The innovation diffusion process in a heterogeneous population: A micromodeling approach. *Management Science*, (36:9), pp. 1057–1079.

Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the Internet topology. Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication, pp. 251-262.

Coleman, J. S. (1958). Relational analysis: The study of social organization with survey methods. *Human Organization*, 17(4):28–36.

Das, K., Schneider, J., and Neill, D. B. (2008). Anomaly pattern detection in categorical datasets. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*. pp. 169–176.

De Reyck, B. and Degraeve, Z. (2003). Broadcast Scheduling for Mobile Advertising. *Operations Research*, 51(4): 509-518.

Goodman, L. A. (1961). Snowball sampling. *The Annals of Mathematical Statistics*, (32:1), pp. 148–170.

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., and Hurst, M. (2007). Patterns of Cascading Behavior in Large Blog Graphs. In *Society of Industrial and Applied Mathematics - Data Mining*, Minneapolis, MN.

McGlohon, M., Akoglu, L., and Faloutsos, C. (2008). Weighted graphs and disconnected components: patterns and a generator. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 524–532. Las Vegas, NV.

McLachlan, G. J. and Peel, D. (2000). *Finite mixture models*. John Wiley and Sons, New York.

Oinas-Kukkonen, H., Lyytinen, K., and Yoo, Y. (2010). Social networks and information systems: Ongoing and future research streams. *Journal of the Association for Information Systems*, 11(2).

Ong, R. (2010). Spamming and mobile marketing: get it right. *International Journal of Intercultural Information Management*, (2:1), pp. 55–67.

Premkumar, G., R. K. and Nilakanta, S. (1994). Implementation of electronic data interchange: an innovation diffusion perspective. *Journal of Management Information Systems - Special section: Strategic and competitive information systems archive*, (11:2).

Pandit, S., Chau, D. H., Wang, S., and Faloutsos, C. (2007). Netprobe: a fast and scalable system for fraud detection in online auction networks. In *Proceedings of the 16th international conference on World Wide Web (WWW '07)*, pp. 201–210.

Salganik, M. J. and Heckathorn, D. D. (2004). Sampling and Estimation in Hidden Population Using respondent-Driven Sampling. *Sociological Methodology*, (34), pp. 193–239.

Teicher, H. (1960). On the mixture of distributions. *The Annals of Mathematical Statistics*, (31:1), pp. 55–73.

Wasserman, S. and Faust, K. (1994). *Social Network analysis: methods and applications*. Cambridge University Press.

Watson, H.J. and Wixom, B. H. (2007). The Current State of Business Intelligence. *Computer*, (40:9), pp. 96–99.

Zhang, B., Zhou, Y., and Faloutsos, C. (2008). Toward a Comprehensive Model in Internet Auction Fraud Detection. In *the Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS '08)*.