On Modeling of Instant Messaging Traffic

Thorsten Kisner and Firoz Kaderali
Department of Communication Systems
Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany
{thorsten.kisner,firoz.kaderali}@fernuni-hagen.de

Abstract

In contrast to the extensive use of Instant Messaging, traffic characteristics have not been the focus of research for a long time, perhaps because of the extremely low bandwidth which they require. When taking the use of IM in mobile or ad-hoc networks into consideration, the importance of bandwidth is an inherent question. Current major IM networks are client/server based with predictable processing and transmission times, but when operating IM on top of a peer-to-peer overlay network, the transmission time (relaying through many nodes) has to be considered. In order to study efficient routing algorithms in ad-hoc or peer-to-peer networks, a realistic user behavior is needed. Although selected aspects of traffic characteristics of IM have been investigated in the past, none of them resulted in a modelling framework. We propose such a framework to model the user behavior in large Instant Messaging networks.

1 Introduction

Instant Messaging (IM) became popular at the end of the 90s and still has a large and ever increasing usage. In a study carried out by IDC it was estimated that approximately $12 \cdot 10^9$ instant messages are sent worldwide each day [1]. While traffic characteristic of IM have not been the focus of research in the past few years, the first aspects of IM or computer-mediated communication ever to be investigated, were social psychological aspects. Long before IM became popular, models like the *reduced social cues* [10] or the *social identity deindividuation model* [14] had been proposed and widely accepted to explain social behaviors with this type of communication.

In section 2 we will present related work, followed by a description of our data records for simulation and directly derived characteristics in section 3. Section 4 will describe additional characteristics for our simulation and the simulation environment itself. We will summarize and evaluate the outcomes in section 5.

2 Related Work

Early studies of IM traffic characteristics consists mainly of limited observation of a small network or interviews with a few users of IM systems. A different social behavior of school and college students was discovered in [8] by interviewing 16 teenagers, revealing a daily usage of mainly below 2 hours and an averaged number of contacted buddies between 2 and 5. About 20 people were interviewed in addition to some log file analysis in [13] to establish that IM is mainly used to check the presence status of a conversational partner and arrange the ongoing discussion in another way (personal meeting or telephone). The conversational data of 8271 messages in 175 conversations were analyzed in [20] concentrating on the coordination and collaboration of tasks in a distributed team which calculated a median conservation length of 6 minutes.

The amount of analyzed data increased in [9] where logfiles of 437 users were analyzed to categorize specific social behavior. A web-based (HTML) chat system is investigated at packet level in [6] and characteristics of TCP session durations, inter-arrival times and packet sizes are explored and compared between the web-based chat system and Internet Relay Chat (IRC) coming to the conclusion of similar packet inter-arrival times, but different distribution of packet sizes.

Full access to traffic information of a large organization led to detailed statistics [21] of about 900 users using the AOL Instant Messenger or the Windows Live Messenger (formerly known as MSN Messenger). As one of the first studies with a lager user group, Smith [16] investigated a XMPP[1] Network with 50.000 users and demonstrated the scale-free property of a buddy[2] network and high clustering coefficient, the local clustering coefficient c_i represents the ratio between the connection of all neighbors of node i to all possible

[1]The IETF Standard Extensible Messaging and Presence Protocol [15] was formerly known as Jabber

[2]Contacts of each IM user are called buddies, the network is created by all users (vertices), an edge from user i to j is created if j is in the buddy list of i.

connections among each other [19]. A high c_i in a social network suggests that people with common friends tend to be friends too.

An impressive study was presented in [11]: the complete MSN network both in size and activity for one month (June 2006) was analyzed. The communication data as well as the presence data and demographic information of nearly $240 \cdot 10^6$ people were collected, resulting in approximately $7 \cdot 10^9$ messages and $64 \cdot 10^6$ users each day. The demographic characteristics the users between 15-35 are strongly overrepresented (compared to the age pyramid of the world population) and that communication is preferably between different genders, but in the same age group and country. The distribution of the shortest path lengths show an average length of 6.6, which is a little bit longer than in the famous study of Travers and Milgram [17] (6 Degrees of Separation) and many articles which followed.

In this paper we present a modular modeling framework for user behavior in Instant Messaging, parametrize the identified modeling functions and compare our results with literature (if available). As far as we know this is the first comprehensive simulation model for IM.

3 Traffic Characteristics

We are faced, like many others, with the lack of comprehensive data of IM usage or log files of IM server. Due to the great similarity between computer-mediated communication using IRC and other chat systems [6], we use amply available IRC data for our modeling. The trace collection for our model is done by capturing four channels on different IRC servers. Table 1 shows an overview of channels, number of users, number of messages and distinct sessions. The number of messages is plotted in figure 1 for all four channels and the capturing time of 94 days, each 24 hours. The peak for the channel `#ubuntu` is based on a new release (7.10) on October 18th 2007. The time-dependence of events are shown in figure 2, all dates are given in GMT+2, thus the minimum of occurrence for all event types of the english spoken channels can be assumed at night for U.S. timezones EST to PST.

Channel	Messages	User	Sessions
#iphone	724.872	9.435	158.109
#debian	402.271	8.310	98.268
#ubuntu	1.076.392	36.983	312.274
#xbox360	392.597	3.353	43.131

Table 1: Overview of captured events

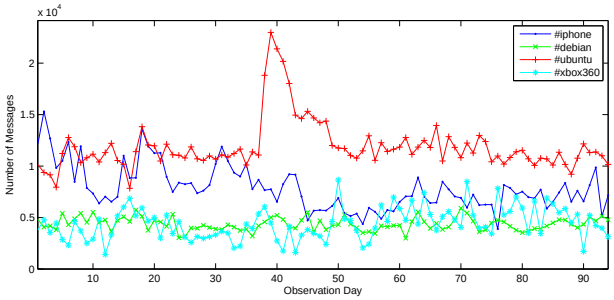


Figure 1: Message Overview

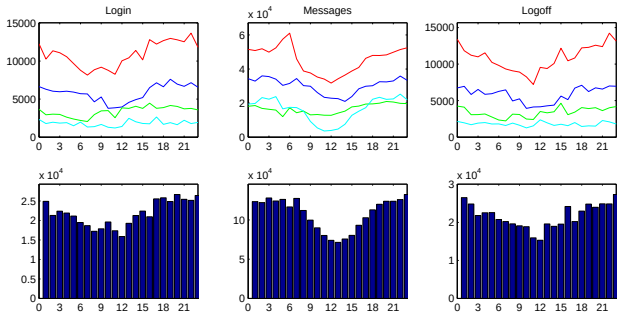


Figure 2: Time-dependence of events

For different characteristics we propose approximations, but renounce using common goodness-of-fit tests like the χ^2 test or Kolmogorov-Smirnov test, because the huge number of samples means that the model hypothesis would be overly rejected too often [5]. Instead of the numerical goodness-of-fit tests we will refer to a QQ-Plot in figure 10.

3.1 Session duration

We have analyzed the login and logoff events and determined more than half a million distinct sessions, which we define as the time between login and logoff. The cdf of the session duration is shown for all four channels in figure 3.

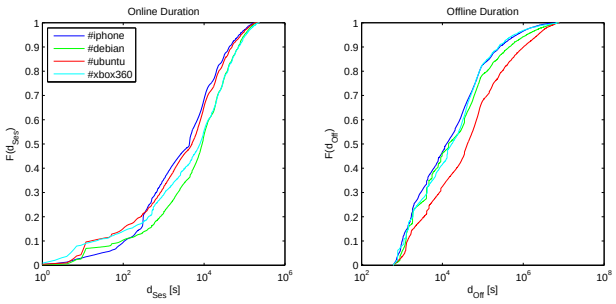


Figure 3: Cdf of session duration

The Weibull distribution was identified in [21] to be best suited to describe the distribution of the session duration. In the generalized form the Weibull distribution is described using three parameters, the scale parameter ϑ, the location parameter τ and the shape parameter α. With $\vartheta > 0$, $\alpha > 0$ and $\tau \in \mathbb{R}$ the cumulative distribution function (cdf) is defined as 0 for $x < \tau$ and otherwise

$$F_{\vartheta,\alpha,\tau}(x) = 1 - e^{-\left(\frac{x-\tau}{\vartheta}\right)^{\alpha}} \quad (1)$$

We have investigated a couple of other distributions (see table 2), which are not described in detail here. Although we approve [21] of the Weibull distribution as a very good fit, we discovered the Dagum is slightly better. Table 2 shows the sum of the quadratic error for each distribution, normalized to the mean error of the best fit.

The Dagum distribution is parametrized by $\beta, \vartheta, \delta > 0$ and [7] proved special characteristics like UBT (upside-down bathtub curve) or DFR (decreasing failure rate) for certain relations of these parameters characterizing this distribution as a very flexible hazard function.

$$F_{\beta,\vartheta,\delta}(x) = \begin{cases} 0 & \text{für } x \leq 0 \\ \frac{1}{\left(1+\frac{\vartheta}{x^{\delta}}\right)^{\beta}} & \text{für } x > 0 \end{cases} \quad (2)$$

	e_1	e_2	e_3	e_4	$\overline{e}$
Weibull	1,11	0,97	1,31	0,65	1,01
Exponential	4,61	3,83	6,27	2,04	4,19
Hjorth	2,52	2,2	3,41	1,52	2,41
Burr	16,47	19,7	24,42	9,58	17,54
Logistic	6,22	6,27	8,58	3,09	6,04
Dagum	1,08	1,07	1,42	0,43	1
Extr. Value	5,5	5,26	7,58	2,64	5,24
Laplace	4,66	4,86	6,33	2,4	4,56
Lognormal	2,63	3,05	4,08	2,35	3,03

Table 2: Goodness-of-Fit for session duration

These calculations for all listed distributions are implicitly done for every further fit, but not mentioned in the text anymore.

The churn rate is implicitly given by the length of the session, reconnection will be available after an offline duration which can be described with a Weibull distribution, having a marginal better error than the the Dagum distribution ($\overline{e} = 1.0015$).

3.2 Message Length

In contrast to [11] where no access to the messages was possible or [21], where only the total message sizes, including HTML formatting or encoding information are considered, we had access to the content of all messages and the exact text length.

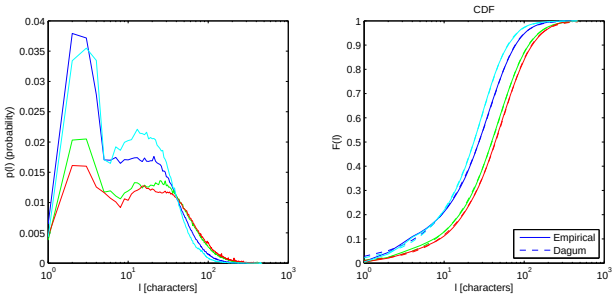


Figure 4: Distribution of Message Length

We studied the message length of about $2.5 \cdot 10^6$ unique messages to derive a suitable distribution for it. Figure 4 shows the probability distribution on the left and the corresponding cdf on the right side. This cdf is similar to [21], but there is overhead information like message header and HTML formatting included and only the total number of bytes considered.

The peaks seen in the probability distribution are remarkable, but this effect can be explained by the excessive use of abbreviations or smileys. Table 3 summarizes the Top 5 messages with such a short length to give an impression of the most used abbreviations.

	l=2	l=3	l=4
1.	ok (7791)	lol (22704)	yeah (4707)
2.	:) (5660)	yes (7300)	haha (3478)
3.	hi (5575)	heh (3043)	hehe (2305)
4.	no (4938)	hmm (2668)	nope (1660)
5.	:p (3083)	yea (1846)	nice (1089)

Table 3: Top 5 Messages

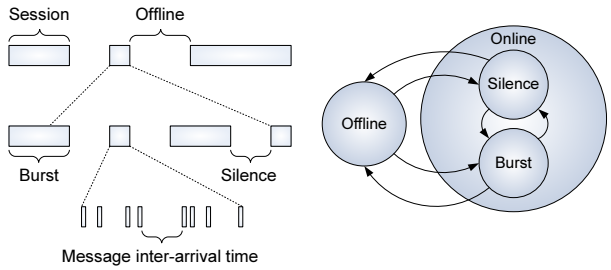
The Dagum distribution (see the dashed line in figure 4 on the right) seems to be a very good fit for the message length[3], apart from outliers, which have a very low message length both curves are nearly identical. These message lengths seem to be dominant, but remember (especially on the left of figure 4) the logarithmic scale of the x-axis.

The is in-line with early studies of IM like [13], where inaccurate grammar and spelling as well as a high usage of abbreviations and very short messages was detected.

[3] Just as for the session duration in table 2 we calculated the best-fit for all other distributions and found the Weibull ($\overline{e} = 2.47$) and the Exponential distribution ($\overline{e} = 2.76$) as the next best ones.

4 Modeling IM Traffic

In order to simulate the behavior of a user we consider in the highest level the *Offline* and *Online* presences. As shown in figure 5a, each *Online* period can be divided into alternating *Burst* and *Silence* periods (will be described in section 4.1).



(a) Specification of times (b) State transitions

Figure 5: Simulation Overview

The possible state transitions are shown in figure 5b. Beginning *Offline* the user can login into the IM network and transits to *Online* by either entering the sub-state *Burst* or *Silence*. Messages can only be sent during *Burst* periods with specific inter-arrival times.

4.1 On/Off Model

To describe systems with alternating active and silence phases an On/Off Model[4] is often used. The On/Off Model is used in various fields of research, e.g. for modeling VBR traffic in ATM networks [12] or describing effects in inter-domain routing protocols [22].

The resolution in time is defined as $\delta = 60$ seconds and we generated time series where each entry represents the number of messages in a period of 1 minute. To minimize the effect of outliers and a disadvantageously choice of δ, we allow several exceptions. If two long *On* Periods are interrupted by a very short *Off* period (depending on the length of the surrounding *On* periods), the *On-Off-On* Sequence will merge to just one *On* period.

The cumulative distribution function (cdf) of both periods are shown in figure 6. We see that 95% of all conservations which are not interrupted by a silence phase are shorter than 10 minutes. While the distribution for all channels are similar in the *On* period, they are rather apart in the *Off* period, but have a similar characteristic.

On Phase The burst period (figure 6 on the left) can be described very well using the Hjorth-

[4] Also known as Burst/Silence model

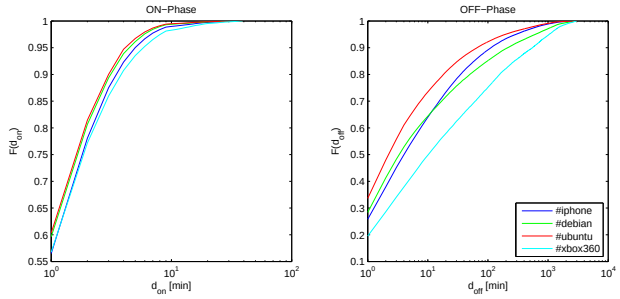


Figure 6: Cdf of retention times (*On* and *Off* period)

Distribution. This distribution is defined for positive coefficients $\alpha, \beta, \vartheta > 0$ and $x \geq 0$ as

$$F_{\alpha,\beta,\vartheta}(x) = 1 - \frac{e^{\frac{-\alpha x^2}{2}}}{(1+\beta x)^{\frac{\vartheta}{\beta}}} \tag{3}$$

and otherwise as zero.

Off Phase The silence period (figure 6 on the right) cannot be easily approximated with a common distribution. We figured out also the Hjorth distribution, but have to accept a sub-optimal approximation for low values (see figure 10).

Distribution of Messages The distribution of the inter-arrival time d_{iat} of messages is shown in figure 7, each (red) line represents the distribution for a specific duration of the *On* period. For the approximation method we use the duration independent distribution (blue line). As mentioned before we set the resolution time for the On/Off Model to 1 minute, thus one could expect a maximum inter-arrival time of 60 seconds, otherwise the system should transit to the Off state. The longer inter-arrival times can be found in the exceptions we described previously and contain about 15% of all inter-arrival times. The Weibull distribution can be applied when describing inter-arrival times of messages during the burst period.

4.2 Topology of Social Network

So far our model accounts all timing aspects for sending a message, but accurate simulations require explicit source and target information ('who talks with whom?'). These relationships create the social network, mentioned before as a buddy network. To simulate the message exchange between distinct users, we have to account for this, but even for a relative small number of users a topology cannot be created manually, thus we investigated specialized graph generation algorithms which are able to reproduce the properties found in real buddy networks.

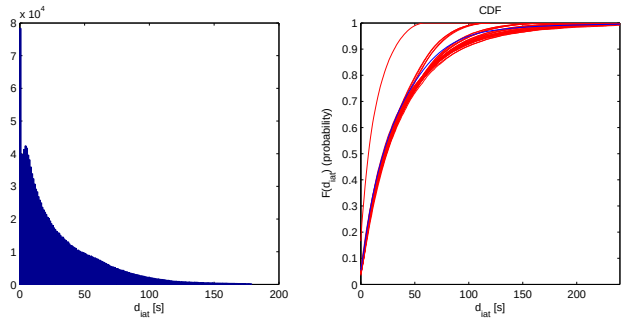


Figure 7: Message inter-arrival time during *On* period

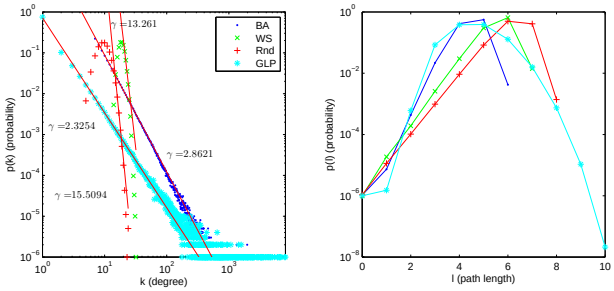


Figure 8: Degree distr. and shortest paths lengths

Such a network has been identified as a scale-free network [3] where some nodes are highly connected, but most are lowly connected. This is independent of the number of nodes n and can be expressed as a power law relationship of the degree distribution $P(k) \sim k^{-\gamma}$. The identified parameters range from $\gamma = 1.8$ [16] to a power-law with exponential cutoff [11] $P(k) \sim k^{-\gamma}e^{-\beta k}$ with $\gamma = 0.8$ and $\beta = 0.03$.

Graph generation algorithms cover a wide range from random, preferential attachment or optimization-based generators up to geographical models or internet specific generators. For reasons of comparison we investigated a random generator and three preferential attachment models which are able to generate power law degree distributions.

The algorithm of Barabási and Albert [2] is used because of its simplicity, but with the certainty that it will create scale-free networks with $\gamma \approx 3$. A modification of the original BA algorithm includes a parameter κ which specifies the number of edges that are generated for each new node, helping to influence the average path length, which increases with $\frac{logn}{loglogn}$ otherwise. We also evaluated the Watts-Strogatz [19] algorithm, which is able to create networks with a clustering coefficient independent of the number of nodes n. Finally we investigated the *Generalized Linear Preference* (GLP) algorithm [4] which can produce degree distributions with $\gamma \in (2, \infty)$ and different clustering coefficients.

Figure 8 shows the generated degree distribution of the social network on the left. Measurements show a global clustering coefficient $\overline{c} = \frac{1}{n}\sum_{i=1}^{n} c_i$ between $\overline{c} = 0.137$ [11][5] and $\overline{c} = 0.33$ [16] while we can produce $\overline{c}$ between 10^{-4} and 0.76 in our simulation environment.

To approximate the distance distribution we calculated the distance between approximately $5 \cdot 10^6$ randomly chosen node pairs for each algorithm. Figure 8 the distribution of the shortest path lengths of the synthetically generated networks on the right. While the coefficient γ of the degree distribution is far away from the measured parameter, we identified the Watts-Strogatz algorithm as a suitable graph generation algorithm, because the average path length $\overline{l}$ as well as the average degree $\overline{d}$ and clustering coefficient $\overline{c}$ are close to real networks [9, 11, 16, 21] (table 4).

	γ	$\overline{l}$	$\overline{c}$	$\overline{d}$
BA	-2,8621	4,5434	0,0002	13,9999
WS	-13,261	5,6463	0,3375	19,0362
Rnd	-15,5094	6,3061	10^{-4}	10
GLP	-2,3254	4,6097	0,7567	2,8451

Table 4: Topology metrics for n=10^6

4.3 Simulation

Now we can put all components together into a simulation environment. The created social network represents the possible contacts (thus possible conversation partners) of all nodes. The period, a user is online in the instant messaging network can be described by the Dagum distribution ($\beta = 0.19983$, $\vartheta = 26301741.8618$, $\delta = 1.6132$). The offline duration is extracted from the periods between sessions and can be described with a Weibull distribution ($\vartheta = 49626.7165$, $\alpha = 0.39129$, $\tau = 948.7647$). The message length of each message can also be described by a Dagum distribution ($\beta = 0.40621$, $\vartheta = 40152.6214$, $\delta = 2.5537$). The occurrence within an online-period can be described with an On/Off model (On: Hjorth ($\alpha = 0.20426$, $\beta = 2.1668 \cdot 10^{-8}$, $\theta = 0.12273$); Off: Hjorth ($\alpha = 4.27 \cdot 10^{-7}$, $\beta = 0.36061$, $\theta = 0.20927$)) where the inter-arrival times are modelled with a Weibull distribution ($\vartheta = 33.136$, $\alpha = 0.94435$, $\tau = -0.50169$).

The social network defines the starting- and endpoint of a conversation. The underlying message routing (client/server architecture, distributed client/server architecture or a peer-to-peer network) can be chosen in the simulation environment. Ex-

[5]The clustering coefficient represents the clustering of people with whom a person communications, the $\overline{c}$ of the buddy network is expected to be higher.

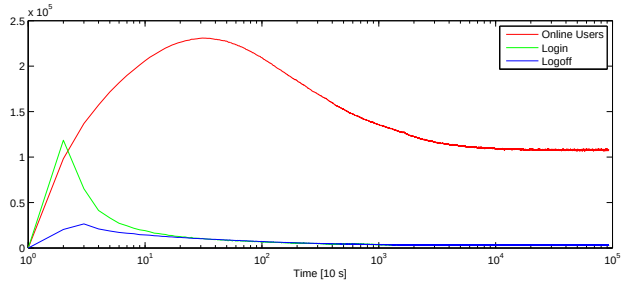


Figure 9: Simulation for $n_{max} = 10^6$ nodes

amples we present here are based on a distributed client/server architecture which can be interpreted as XMPP. We assume an amount of 10000 XMPP server and each node N_i belongs to a randomly chosen Server S_j. Figure 9 shows the number of users who are online and the number of login and logoff events per time interval. The system leads to a balance of login and logoff events after approximately 8 hours. Keeping in mind the time-dependence of login and logoff events (figure 2) we can add this characteristic to the probability of each event. This will simulate the dependence of time of day for users in specific geographic areas.

Time	N_{src}	S_{src}	S_{dst}	N_{dst}	Length
71378.7	891	27	375	886	338
71397.2	891	27	375	886	71
71437.0	891	27	497	887	13
71441.1	891	27	497	887	12
71456.3	891	27	497	887	3

Table 5: Example of generated traffic

The simulation tool creates trace files in XML format, which can be easily transformed to any desired format. Table 5 shows a short cut-out of the generated traffic for a randomly chosen node.

5 Conclusion and Future Work

By plotting the empirical quantil against the expected quantil (derived from our proposed distributions) in a QQ-Diagram (figure 10, each distribution with the scale normalized to 1) we can evaluate the goodness-of-fit, a linear plot indicates a perfect fit. Beside the duration of the *Off* period all approximations show really good fits of the expected data. In this case the distribution is not able to produce a rapid slope for small values, followed by a smoothly slope.

We identified the Dagum distribution to be slightly better at describing the session duration than the formerly identified Weibull distribution [21].

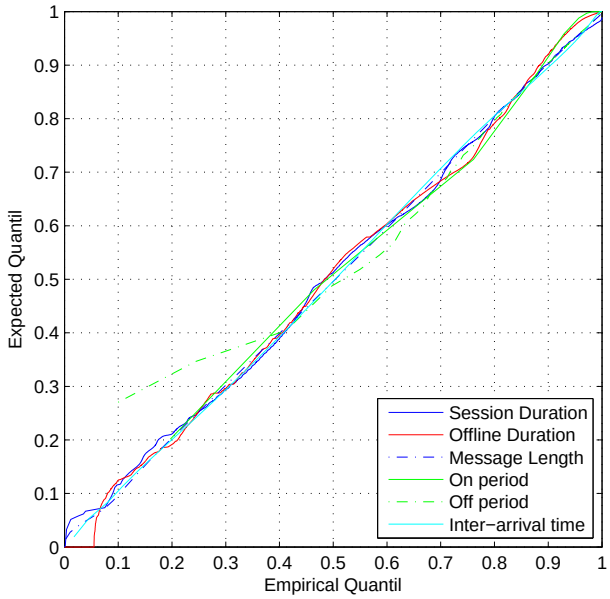


Figure 10: Quantil-Quantil Plot

As a novelty we characterize the inter-arrival time and length of messages with very good approximations. Our model is implemented into a event-based simulation framework which can generate trace files for other simulation environments or directly simulate the user behavior. Currently we are preparing the release of our Java based simulation environment as open source software. In this initial version users act stochastically independent, future work will be include feedback loops to describe the behavior more realistic, e.g. a user will react (answer) if he got a message. Another improvement will be the provision for different presences like *Away* or *Do not Disturb* and the propagation of changes of them.

The measured average number of messages per (active) user per day is 109 [11], our simulation environment generates, after the balanced state, approximately 117 messages per user per day with 70,2% of active users (compared to 68,8% in [11]).

Using a topology with scale-free and small-world properties we are able to describe the social network of an instant messaging system in a massive graph with 10^6 nodes, knowing that the generation algorithm has some flaws to reproduce structures found in real instant messaging networks, especially the degree distribution. But the network topology is – like the other solutions – one component in the modelling framework which can be easily exchanged. Thus we accepted this compromise, but identified this issue for further optimization, more general algorithms to generate scale-free graphs (e.g. with $\gamma \in (1, \infty)$ [18]) will be investigated for their suitability of modelling IM networks.

References

[1] *Worldwide Enterprise Instant Messaging Applications 2005-2009 Forecast and 2004 Vendor Shares: Clearing the Decks for Substantial Growth*. IDC Market Analysis, 2005.

[2] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 288:509–512, 1999.

[3] A. L. Barabási and E. Bonabeau. Scale-free networks. *Scientific American*, 288(5):60–69, May 2003.

[4] Tian Bu and Don Towsley. On distinguishing between internet power law topology generators. In *Proceedings of the Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2002)*, volume 2, pages 638–647, 2002.

[5] Ralph B. D'Agostino and Michael A. Stephens. *Goodness-Of-Fit-Techniques*. Marcel Dekker Ltd, 1986.

[6] Christian Dewes, Arne Wichmann, and Anja Feldmann. An analysis of internet chat systems. In *IMC '03: Proceedings of the 3rd ACM SIGCOMM conference on Internet measurement*, pages 51–64, New York, NY, USA, 2003. ACM.

[7] Filippo Domma and Mariangela Zenga. The dagum distribution as survival model. In *28th Annual Conference of the International Society for Clinical Biostatistics*, Athens, Greece, 2007. Myriki.

[8] Rebecca E. Grinter and Leysia Palen. Instant messaging in teen life. In *CSCW '02: Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 21–30, New York, NY, USA, 2002. ACM Press.

[9] Ellen Isaacs, Alan Walendowski, Steve Whittaker, Diane J. Schiano, and Candace Kamm. The character, functions, and styles of instant messaging in the workplace. In *CSCW '02: Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pages 11–20, New York, NY, USA, 2002. ACM.

[10] Sara Kiesler, Jane Siegel, and Timothy W. McGuire. Social psychological aspects of computer-mediated communication. In *Computer-supported cooperative work: a book of readings*, pages 657–682. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[11] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of 17th International World Wide Web Conference (WWW2008)*, Beijing, China, 2008.

[12] R. Manivasakan, U.B. Desai, and Abhay Karandikar. End-to-end simulation of vbr traffic over atm networks using cipp network traffic model. *Third International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'99)*, 00:338, 1999.

[13] Bonnie A. Nardi, Steve Whittaker, and Erin Bradner. Interaction and outeraction: instant messaging in action. In *CSCW '00: Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 79–88, New York, NY, USA, 2000. ACM.

[14] S. D. Reicher, R. Spears, and T. Postmes. A social identity model of deindividuation phenomena. *European Review of Social Psychology*, 6:161–198, 1995.

[15] P. Saint-Andre. Extensible Messaging and Presence Protocol (XMPP): Core. RFC 3920 (Proposed Standard), October 2004.

[16] Reginald D. Smith. Instant messaging as a scale-free network, 2002.

[17] Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.

[18] D. Volchenkov and Ph Blanchard. An algorithm generating scale free graphs, 2002.

[19] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[20] Stephanie Woerner, JoAnne Yates, and Wanda Orlikowski. Conversational coherence in instant messaging and getting work done. *40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pages 77–87, IM 2007.

[21] Zhen Xiao, Lei Guo, and John Tracey. Understanding instant messaging traffic characteristics. In *ICDCS '07: Proceedings of the 27th International Conference on Distributed Computing Systems*, page 51, Washington, DC, USA, 2007. IEEE Computer Society.

[22] Xiaoliang Zhao, Dan Massey, Mohit Lad, and Lixia Zhan. On/off model: A new tool to understand bgp update burst. Technical report, University of Southern California, 2004.